دليل أبحاث حوسبة اللغة العربية

مطبعة عمال المطابع
Printers Press

اللجنة الوطنية الأردنية للنهوض
باللغة العربية للتوجه نحو مجتمع المعرفة

دليل أبحاث حوسبة اللغة العربية

الجزء الثاني

إعداد

فريـــق عمـــل

دليل أبحاث حوسبة اللغة العربية

ربيع الأول ١٤٤٠هـ/ كانون الأول ٢٠١٨م


بسم الله الرحمن الرحيم

اللجنة الوطنية الأردنية للنهوض باللغة العربية للتوجه نحو مجتمع المعرفة

الأستاذ الدكتور

خالـد الكـركي

رئيسًا

الأعضـــاء

الأستاذ الدكتور
إسحق فرحان

الأستاذ الدكتور
كامل العجلوني

الأستاذ الدكتور
عبداللطيف عربيات

الأستاذ الدكتور
محمد حمدان

الأستاذ الدكتور
أحمد هليل

الأستاذ الدكتور
محمد عدنان البخيت

الأستاذ الدكتور
فاروق صبحي زيد الكيلاني

الأستاذ الدكتور
همام غصيب

أمين عام وزارة التربية والتعليم – مندوبًا عن الوزارة (الأستاذ محمود العكور)

أمين سر اللجنة الوطنية الأردنية للتربية والثقافة والعلوم (إبتسام أيوب)

فريق العمل

الأستاذ الدكتور
محمد زكي خضر
رئيسًا

الأستاذ الدكتور
محمد السعودي
(أمين عام مجمع اللغة العربية) عضوًا

الدكتور
مجدي صوالحة
(الجامعة الأردنية) عضوًا

الدكتور
سامي عبابنة
(الجامعة الأردنية) عضوًا

الدكتور
يوسف حمدان
(الجامعة الأردنية) عضوًا

الأستاذ
مأمون الحطاب
(شركة النص العربي) عضوًا

ملخص الدراسة

تروم هذه الدراسة بالدرجة الأولى إطلاع الباحثين العرب - اللغويين خاصة - على الجهود البحثية المكتوبة باللغة الإنجليزية والمبذولة في سبيل حوسبة اللغة العربية، وذلك بهدف ردم الهوة بين مجالي عمل الحاسوبيين واللسانيين وتوحيد جهودهما، إضافة إلى تلخيص لما نشر باللغة العربية حول الموضوع.

وقد جاءت الدراسة في مقدمة وأربعة فصول وملحق بالمصطلحات، كشفت المقدمة عن أهمية حوسبة اللغة العربية، وما يصاحبها من إشكاليات، وآفاق حلّها، ثم توضيح أهداف الدراسة ومنهجيتها وآلية التنفيذ.

وقدم الفصل الثاني دراسة مسحية لأبرز الدراسات العربية في مجال حوسبة اللغة العربية، وقد تُتبعت الجهود التي بذلت في سبيل تحقيق هذه الغاية، ونُظر إلى هذه الدراسات بالالتفات إلى بدايات تشكل الوعي بأهمية هذه القضية فعُرضت أهم الدراسات الشمولية التي سعت إلى إبراز أهمية الحوسبة للغة العربية، ثم أُظهر اهتمام الباحثين العرب بمحاولة حوسبة قواعدية اللغة في جانبي النحو والصرف، ثم أُتبع ذلك بمبحثين تناول الأول ما يُعرف بالمصادر. أما المبحث الثاني فاهتم بالأبحاث المعنية بالتطبيقات الحاسوبية ذات الصلة باللغة العربية.

وقدم الفصل الثالث ترجمة لأهم ما نشر من أبحاث ودراسات ومؤلفات عن المعالجة الحاسوبية للغة العربية في المجلات العلمية المحكمة المتخصصة والمؤتمرات العلمية وما يتوفر على الشابكة (الإنترنيت) وتبويبها حسب التخصصات الفرعية، وقد تمت العملية باختيار (٥) أوراق من الأبحاث المسحية المهمة باللغة الإنكليزية ووضع خلاصات لها بحدود ٢٠٠٠ كلمة لكل بحث، واختيار (٤٥) بحثًا من الأبحاث المهمة باللغة الإنكليزية ووضع خلاصة طويلة لها باللغة العربية بحدود ١٠٠٠ كلمة للبحث الواحد، ثم اختيار (١٠٠) بحث

باللغة الإنكليزية ووضع خلاصة مختصرة لها باللغة العربية بحدود ٤٠٠ كلمة لكل بحث. وقد وزعت الأبحاث على ثلاثة جوانب : القواعدية اللغوية التي حصرت في جانبي النحو الصرف، ثم قضايا الحوسبة ومستلزماتها في جانب المصادر وشملت خمسة مواضيع فرعية هي: المعاجم الآلية، والمكنز، والمدونات الموسومة، والأنطلوجيا، وشبكات الكلمات. أما التطبيقات الحاسوبية، فضمت جميع المباحث التأسيسية الضرورية تنظيرًا وتطبيقًا: توزعت التطبيقات على موضوعات فرعية عددها ١١ موضوعاً، هي: التعرف الصوتي وتوليد الكلام، والقارئ الآلي، والترجمة الآلية، والتحليل الدلالي، والسؤال والجواب، وتلخيص النصوص، وتحليل الرأي، والتعرف على أسماء الأشياء، والتعليم والتعلم الآلي. وقد ارتئي في اختيار الأبحاث أن تكون في غالبيتها مما نشر في السنوات الأخيرة لكي تعطي فكرة عن آخر ما توصل إليه البحث العلمي في حقول موضوعاتها. وقد وضع الفصل الثالث هذا في جزء ثان مستقل

وقد أبرز الفصل الرابع أهم النتائج والتوصيات التي خلصت إليها الدراسة...

واستلزمت الدراسة الوقوف عند المصطلحات الكثيرة التي تواجه الباحث في هذا المجال فتناول الملحق في نهاية الدراسة أكثر من (٧٠٠) مصطلح في مجال حوسبة اللغة الطبيعية مما تتطلبه دراسات من هذا القبيل؛ مما يؤمل أن يسهم في حلّ كثير من الإشكاليات التي تواجه الدارسين خاصة من اللغويين بإطلاعهم على هذه المصطلحات ومفاهيمها. وقد تضمن الجدول الذي ضم تلك المصطلحات تعريفا لها.

شكر وعرفان

يسر فريق العمل أن يتقدم بالشكر والعرفان "للجنة الوطنية للنهوض باللغة العربية للتوجه نحو مجتمع المعرفة" لتبنيها هذه الدراسة، ولرئيس اللجنة وأعضائها لتوجيهاتهم السديدة.

ويقدر فريق العمل جهود المساعدين كافة، الذين أسهموا في المساعدة في تلخيص الأبحاث وترجمة خلاصاتها، ومعالجتها حاسوبيًا، والوصول بها إلى صورتها النهائية.

قائمة المحتويات

الموضوع	الصفحة
ملخص الدراسة	٥
شكر وعرفان	٧
قائمة المحتويات	٩
الفصل الأول التمهيد	١١
١.١ أهمية الدّراسة وأهدافها	١٣
١.٢ مجتمع الدراسة	١٣
١.٣ آلية التنفيذ	١٤
١.٤ مدة إنجاز العمل ومراحلها	١٥
صعوبات الدراسة ومحدداتها	١٥
١.٦ مقدمة عن الحوسبة	١٨
١.٧ حوسبة اللغة	١٩
١.٨ مناهج البحث والتطوير في حوسبة اللغة	٢٠
١.٩ التحديات التي تواجه العاملين في حوسبة اللغة العربية	٢٣
١.١٠ الموارد الأساسية في عملية حوسبة اللغة العربية	٢٤
١.١١ محتويات التقرير	٢٦
الفصل الثاني واقع الدراسات العربية في حوسبة اللغة العربية	٢٧

٢- ١- مقدمة ٢٩

٢-٣-١- الحوسبة والمستويات اللغوية الأساسية للغة العربية ٣١

٢-٣-٢- المستوى النحوي ٤٧

٢-٣-٣- المستوى الصرفي ٥٦

٢-٣-٤- الموارد ٦٦

٢-٣-٥- التطبيقات ٨٨

٢-٤- خاتمة ١١١

الفصل الثالث خلاصات ترجمة الأبحاث الإنكليزية ١١٥

٣-١ تمهيد ١١٦

٣-٢-١ أبحاث الصرف ١١٧

٣-٢-٢ أبحاث النحو ١٥٨

٣-٣ أبحاث الموارد ١٨٢

٣-٣-١ أبحاث المكنز ١٨٣

٣-٣-٢ أبحاث المدونات الموسومة ٢٠٥

٣-٣-٣ أبحاث المعاجم الآلية ٢٢٢

٣-١-٤ أبحاث الأنطولوجيا ٢٤٤

٣-٣-٥ أبحاث شبكات الكلمات ٢٦٩

٣-٤ أبحاث التطبيقات ٢٨٧

٣-٤-١ أبحاث التعرف الصوتي وتوليد الكلام ٢٨٨
٣-٤-٢ أبحاث القارئ الآلي ٣٣٢
٣-٤-٣ أبحاث التعرف على أسماء الأشياء ٣٥٠
٣-٤-٤ أبحاث التشكيل الآلي ٣٩٦
٣-٤-٥ أبحاث البحث في النصوص ٤٢٠
٣-٤-٦ أبحاث الترجمة الآلية ٤٣٢
٣-٤-٧ أبحاث السؤال والجواب ٤٥٥
٣-٤-٨ أبحاث تلخيص النصوص ٤٧٧
٣-٤-٩ أبحاث التحليل الدلالي ٥٠٣
٣-٤-١٠ أبحاث تحليل الرأي ٥٢٠
٣-٤-١١ أبحاث التعليم والتعلم الآلي ٥٤٧
الفصل الرابع: المناقشات والتوصيات ٥٦١
الملحق ١: المصطلحات ٥٧١

الفصل الثالث

خلاصات ترجمة الأبحاث الإنكليزية

٣-١ تمهيد

٣-٢ أبحاث اللغة

٣-٢-١ أبحاث الصرف

٣-٢-٢ أبحاث النحو

٣-٣ أبحاث الموارد

٣-٣-١ أبحاث المعاجم الآلية

٣-٣-٢ أبحاث المكنز

٣-٣-٣ أبحاث المدونات الموسومة

٣-٣-٤ أبحاث الأنطولوجيا

٣-٣-٥ أبحاث شبكات الكلمات

٣-٤ أبحاث التطبيقات

٣-٤-١ أبحاث التعرف الصوتي وتوليد الكلام

٣-٤-٢ أبحاث القارئ الآلي

٣-٤-٣ أبحاث التعرف على أسماء الأشياء

٣-٤-٤ أبحاث التشكيل الآلي

٣-٤-٥ أبحاث البحث عن النصوص

٣-٤-٦ أبحاث الترجمة الآلية

٣-٤-٧ أبحاث السؤال والجواب

٣-٤-٨ أبحاث تلخيص النصوص

٣-٤-٩ أبحاث التحليل الدلالي

٣-٤-٠١ أبحاث تحليل الرأي

٣-٤-١١ أبحاث التعليم والتعلم الآلي

الفصل الثالث

٣-١ تمهيد

توزعت الأبحاث في هذا الفصل على ثلاث مجموعات هي أبحاث اللغة وأبحاث الموارد وأبحاث التطبيقات.

٣-٢ أبحاث اللغة

تشمل أبحاث اللغة موضوعين هما: الصرف والنحو

٣-٢-١ أبحاث الصرف

وتضم ثلاثة عشر بحثاً، منها خمسة أبحاث من النوع (أ) ويقصد بها المحللات الصرفية المتحكّم بغموضها لنمذجة اللغة العربيّة الحديثة باستخدام شبكات الحالات المنتهية، وتحسين استخراج الجذور باستخدام أدوات التجذيع الخفيف، و«مداميرا» : أداة سريعة وشاملة للتحليل الصرفي وفكّ اللبس في اللغة العربيّة، وتطوير محلل صرفي عربي مرن باستخدام تكنولوجيا الحالة المنتهية، و المحلل الصرفي المعياري للغة العربيّة: «سلمى».

وثمانية أبحاث من النوع (ب)، تتضمن: استخراج الجذور العربيّة باستخدام التحليل الصرفي، ومجموعة الوسم القياسية التي تفسّر الخصائص الصرفية التقليدية للغة العربيّة لوسم أقسام الكلام، والتحليل الحاسوبي الصرفي المستند للمزايا وتوليد لغة عربيّة حديثة، والتجذيع الخفيف لاسترداد المعلومات العربيّة، والمنهجية الجديدة لاستخراج الجذور العربيّة: استخدام العلاقات بين حروف الكلمة ومواقعها فيها لاستخراج الجذر العربي، والتحليل وتوليده الصرفي العربي محدود الحالة، والتحليل الصرفي في بيئة تعليمية للغة العربيّة بمساعدة الحاسوب المسماة «تيلا»، والتحليل الصرفي العربي المعتمد على السياق لغرض الترجمة الآلية.

Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK

Year of publication: *2006*

An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks

Mohammed A, Attia

المحللات الصرفية المتحكم بغموضها لنمذجة اللغة العربيّة الحديثة باستخدام شبكات الحالات المنتهية

الغموض الصرفي مصدر قلق رئيسي للمحللات النحوي (syntactic parsers)، ولبرامج الوسم أقسام الكلام (POS Taggers)، ولباقي أدوات المعالجة الطبيعية. فكلما زاد عدد التحاليل الصرفية للمدخلة المجمعجيّة (lexical entry)، زاد الوقت الذي يحتاجه المحلل النحوي لتحليل الجملة، وزادت عدد التحليلات النحوية الناتجة. يعد المحللان الصرفيان (Xerox & Buckwalter) أفضل المحللات المعروفة والموثَّقة الموجودة للغة العربيّة الحديثة (MSA). ما زال هناك مشاكل في التصميم والتغطية لكلا النظامين مما يزيد من نسبة معدلات الغموض. وتبين هذه الدراسة كيفية بناء محلل صرفي للغة العربيّة متحكّم بالغموض فيه، بُني باستخدام منهجية تعتمد القواعد (rule-based system)، التي تعد جذع الكلمة (stem) الشكل الأساسي لها، وتستخدم هذه المنهجية تقنية الحالة المنتهية (technology finite state).

وأجري التقييم في هذه الدراسة لنظامي (Xerox & Buckwalter) والنظام الجديد وقورنت الأداء وحلّلت.

المقدمة

الغموض الصرفي في العربيّة مسألة معروفة، ولم تجرِ معالجتها بشكل كافٍ، وهذا الغموض يؤدي إلى عقبات في طريق برامج وسم أقسام الكلام (POS Taggers)، والمحللات النحوية (syntactic parses)، والترجمة الآلية (machine translation).

مصادر الغموض الصرفي في العربيّة

العديد من كلمات اللغة العربيّة متجانسة شكلياً (إملائياً) (homographic): تشترك هذه الكلمات بتجانس شكلي، ولكن باختلاف في اللفظ. وهناك عدة عوامل تساهم في هذه المشكلة، منها:

١- عمليات التغيير الإملائي (شكل الكلمة) كالحذف والإدغام (assimilation). مثلاً، الشكل الإملائي "يعد" يحتمل خمسة أوجه؛ يعِد (أعاد)، يعُد (عاد)، يعِد (وعد)، يعُدّ (عدّ)، يُعِدّ (أعدّ).

٢- بعض الكلمات المجردة (lemmas) مختلفة فقط في احتواء أحدهما على حرف مضعّف غير مكتوب. مثلاً، "علم" لها وجهان، عَلِمَ، وعَلَّمَ.

٣- كثير من عمليات التصريف (inflectional operation) تُحدث تغيراً في اللفظ بدون تغير الشكل الإملائي وذلك لغياب الحركات (علامات التشكيل) (diacritics). مثلاً، «أرسل» لها ثلاثة أوجه، أَرسَلَ، أُرسِلَ، أَرسِل.

٤- بعض السوابق واللواحق قد تكون متجانسة. مثلاً، السابقة «ت» تعبر عن الفعل المسند إلى المؤنث الغائب، أو الفعل المسند إلى المذكر المخاطب. فالفعل "تكتب"، يحتمل أن يكون «تَكتُبُ» (أنت)، أو «تَكتُبُ» (هي).

٥- عند إضافة السوابق أو اللواحق للكلمة، من الممكن صدفة أن تجانس كلمة أخرى (coincidental identity). مثلاً، كلمة «أسد»، لها وجهان، «أَسُدُّ» و«أَسَدٌ» (حيوان مفترس).

٦- عند إضافة الزوائد للكلمة، من الممكن صدفة أن تجانس كلمة أخرى. مثلاً، «علمي» لها وجهان؛ «علمي» كلمة واحدة، أو «علمي» مكونة من «علم" والزائدة «ي».

٧- ويوجد تجانس في الكلمات المختلفة بالتصريف، وقد تختلف أو لا تختلف في اللفظ، وعادةً ما تختلف بالمعنى وتنتمي لأقسام كلام مختلفة. مثلاً، كلمة «ذهب» لها وجهان «ذَهَبٌ»و"ذَهَبَ».

الفرضيّة الهرميّة للغموض

تعتبر الفرضية الهرميّة أنّ النظام الغني والمعقد للتصريف والإلصاق في اللغة العربيّة يساعد على تقليل الغموض بدلاً من زيادته. والكلمات المجردة (unmarked stems) تكون غامضة، فإذا صُرِّفت أو أُلصقت بها الزوائد يقلّ غموضها. فمثلا، الكلمة المجردة أو الجذع «كتب»، يمكن تصريفها إلى «يَكتبُ» أو «يُكتَبُ»، ويمكن إلصاق بعض الزواد بها فتصبح «يَكْتُبُه».

استراتيجيات تطوير الصرف العربي.

اللغة العربيّة معروفة بثرائها الصرفي وتعقيدها، ويعد الصرف العربي تحدياً حاسوبياً في المعالجة الحاسوبية وتقنيات التحليل الصرفي. هناك استراتيجيتان لتطوير التحليل الصرفي للغة العربيّة بالاعتماد على مستوى التحليل:

١- الصرف المعتمد على الجذع: الجذع هو الشكل غير المصرّف للكلمة دون السوابق واللواحق والزوائد في بداية الكلمة ونهايتها. والجذع من الأفعال هو الفعل الماضي المسند إلى المخاطب المفرد، ومن الأسماء والصفات يكون الجذع على صيغة النكره المفرد.

٢- الصرف المعتمد على الجذر: يتكوّن الجذر من ثلاثة أحرف (وبشكل قليل من حرفين أو أربعة أحرف) (radicals). ويعرف الوزن للكلمة بأنه عبارة عن قالب من حروف العلة أو الحروف الساكنة، مع وجود أماكن لإضافة أحرف الجذر.

أنظمة المحللات الصرفية العربيّة الموجودة

هناك محللات صرفية متعددة في العربيّة، بعضها متاح للبحث والتقييم، والبقية موجهة للتطبيقات التجارية. مثال على ذلك:

نظام Buckwalter للتحليل الصرفي العربي

وهو نظام يحتوي على ٣٨.٦٠٠ كلمة مجردة. واستخدم في العديد من الأنظمة العالمية (LDC Arabic POS taggers، Penn Arabic Treebank and Prague Arabic Dependency treebank).

ومن عيوب هذا النظام أنه لا يعتمد على القواعد في عمله، ولا يمكن استعماله مولدا للكلمات، وغير مناسب لتحليل التعابير اللغوية متعددة الكلمات (Multi-word expressions (MWEs))، بالإضافة إلى عدد من العيوب الأخرى (راجع البحث الأصلي).

نظام Xerox للتحليل الصرفي

وهو مبني على تقنية الحالة المنتهية (Finite state) وتعتمد على مقاربة الجذر والوزن. ويضم هذا النظام ٤٩٣٠ جذرا و٤٠٠ وزن، وينتج ٩٠٠٠٠ جذع. وله أفضلية لاعتماده على القواعد، وله تغطية واسعة.

ومن عيوب هذا المحلل: الاشتقاقات الزائدة غير المنطقية وغير المستعملة في اللغة، والقصور في تصنيف أقسام الكلام.

وصف النظام

بُني النظام باستخدام تقنية الحالة المنتهية، وهي مناسبة لكل من التحليل والتوليد. ويعتمد النظام على ذخيرة لغوية معاصرة مكونة من ٤.٥ مليون كلمة من مقالات إخبارية، واعتمدت الجذع أساسا للنظام. واحتوت الذخيرة على ٩٧٤١ كلمة مجردة، و٢٨٢٦ تعبيراً متعدد الكلمات. وفّرت بنية النظام تغطية فعّالة للغة العربيّة الحديثة (MSA) في مجال مخصص (المقالات الإخبارية). والنظام متاح للتجربة والبحث العلمي على الموقع (www.attiaspace.com)، مع أدوات الحالة المنتهية المناسبة: محلل معجمي (tokemizer)، ومبسط للفراغات بين الكلمات (whitespace nromalizer)، ومخمن صرفي (morphological guesser).

إحدى نقاط القوة في هذا النظام هي تغطية التعابير متعددة الكلمات، وبفعالية عالية يمكن للنظام التعرف على الأسماء المركّبة للأشخاص والأسماء والمؤسسات، بالإضافة للعديد من التعابير المعقّدة التي تخضع لتغيّرات معجمية وصرفية.

وأحد عيوب هذا النظام هو التغطية المحدودة، كما لا يستطيع النظام التعامل مع النصوص

المشكولة (diacritized texts). ومن العيوب الأخرى أنّ النظام لايستطيع إعادة الحركات (التشكيل) إلى الكلمات، كما لا يحتوي النظام على قوائم للكلمات باللغة الإنجليزية.

تقنية الحالة المنتهية

نجحت هذه التقنية في تطوير المحللات الصرفية للعديد من اللغات بالإضافة إلى اللغات السامية، ويوجد عدد من المزايا لاستخدام هذه التقنية:

- قدرتها على التعامل مع كلّ من الصرف الإلصاقي وغير الإلصاقي.
- سريعة وفعالة، تستطيع التعامل مع شبكة الحالات المنتهية مكونة ضخمة لتمثيل معاجم للكلمات وتصريفاتها.

دعم اللغات التي لا تُكتب بحروف لاتينية (كالعربية) من خلال دعمها لنظام ترميز الحروف الدولي الموحد (unicode).

أداة الحالة المنتهية Xerox تعمل ضمن جميع أنظمة التشغيل.

وهي انعكاسية بشكل كامل، حيث يمكن استعمالها في تحليل النصوص وتوليدها.

القواعد والتعابير المنتظمة (regular expression) المستخدمة في تصميم شبكات الحالات المنتهية، مقروءة وواضحة لتشابهها مع الصيغ اللغوية المعيارية.

وفي النهاية حققنا محولا (transducer) ذا علاقة ثنائية بين مجموعتين من السلاسل الرمزية (strings)، الأولى تحتوي على الكلمة المدخلة (surface form) وتعرف باللغة الدنيا، والثانية تحتوي النماذج المعجمية (lexical forms) أو التحليل وتعرف باللغة العليا. المثال التالي يوضح اللغة الدنيا واللغة العليا للفعل "يشكرون".

Upper Language: شكر + masculine + present + plural + 3rdPerson Lower Language: يشكرون

معالجة التصريف (Morphotactics) في اللغة العربيّة

ويقصد بالتصريف هو دراسة تكوين الكلمات من مكوّناتها (الأجزاء الصرفية)؛ فإما أن تكون إلصاقية بإضافة السوابق واللواحق إلى جذع الكلمة، أو غير إلصاقية وذالك بحدوث تغيّرات على جذوع الكلمات لتحمل المعلومات الصرفية والنحوية.

التقنيات المستخدمة في تحديد حالات الغموض

بهدف جعل النظام ينتج الحلول المقبولة، ويتجاهل الحلول غير المقبولة، يتبع النظام الاعتبارات والتقنيات الآتية:

- استخدام الجذع نموذجا أساسيا، تحاشيا لاشتقاق كلمات غير مستخدمة في اللغة عند اعتماد الجذر أساسا للنظام.
- عدم إضافة كلمات عربيّة كلاسيكية، أو المعاني المختلفة للكلمة، لأنها تزيد من حجم المعجم ومن مستوى الغموض.
- رصد القواعد التي تحكم دمج الكلمات مع اللواصق والإضافات، أو الخصائص اللغوية والمعجمية، التي تعمل مرشحاتٍ للتحليلات غير الصحيحة والغامضة.
- تحديد أي أفعال لها صيغ المبني للمجهول؛ من ١٢٩٧ فعلا، كان فقط (٤١٪) ٥٤٤ فعلا يقبل الإسناد لصيغة المبن ي للمجهول (وهي ٥٠٠ فعل متعد، و٤٤ فعلا لازما). بشكل ابتدائي، جميع الأفعال المتعدية تقبل الإسناد لصيغة المبني للمجهول، أما الأفعال اللازمة فلا تقبل الإسناد إلى المني للمجهول.
- تحديد أيّ الأفعال لها صيغة الأمر؛ من ١٢٩٧ فعلا، كان فقط (٣٧٪) ٤٨٣ فعلا يقبل صيغة الأمر (وهي ٣٤٢ فعلا متعديا، و١٥٩ فعلا لازما).

التقييم

هدفنا تقييم كلّ من المحللات الصرفية (Xerox, Buckwalter and Attia) مع مراعاة الغموض. بسبب عدم وجود ذخيرة عربيّة معيارية وموسومة، فإنه من غير الممكن إجراء

عملية التقييم بشكل واسع وشامل وآلي. لذلك، حضرت تجربة تقييم يدوية محدودة لفحص معدلات الغموض في المحللات الصرفية الثلاثة. اختيرت خمسة مقالات حديثة احتوت على ٩٥٠ كلمة مختلفة، و٦٧ تعبير متعدد الكلمات. وتم فحص جميع الكلمات بواسطة كلّ محلل صرفي على حدة، فكانت نتيجة الدقة في (Buckwalter) ٦٤٪، وفي (Attia) ٧٩٪.

أما الغموض في الإنجليزية فقد فُحص باستخدام محلل XLE الصرفي على ٩٧٩ كلمة لاستخدامها بياناتِ أساسٍ (baseline)، فنتج بعد تحليلها ١٧٣٢ حلاً، بنسبة غموض ١.٧٦٪.

وكانت نسبة دقة الغموض في هذه المحللات على النحو الآتي:

- Xerox : ٤.٣٢٪
- Attia : ١.٧٥٪
- Buckwalter : ٢.٦٪
- XLE English : ١.٧٦٪

وأشارت تقارير الخطأ إلى أن مصادر الغموض في المحللين (Xerox and Buckwalter) تتلخص بما يأتي:

- ضم كلمات عربيّة كلاسيكية.
- عدم تكامل العلائقيّة المعجمية والنحوية.
- تطبيق غير صحيح لقواعد التهجئة.

التوصيات

إن غنى الصرف وتعقيده في اللغة العربيّة، لا يعني أنها تحتوي على غموض كبير، فالتحليل والتقييم أظهرا أن معظم الحالات الغامضة التي نتجت جراء استخدام Buckwalter & Xerox هي زائفة بسبب إدخال الكلمات الكلاسيكية، والجذوع التي ليس لها استخدام في اللغة، وقواعد

تمثيل الكلمات، والخصائص اللغوية، وخصائص التهجئة، ومع تجنب هذه الهفوات يمكن إنتاج محلل صرفي بدرجة غموض أقل.

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 157 -166, Shanghai, China, October 30 - November 1, 2015

Year of publication: 2015

Enhancing Root Extractors Using Light Stemmers

Mahmoud El-Defrawy, Yasser El-Sonbaty, Nahla A. Belal

تحسين استخراج الجذور باستخدام أدوات التجذيع الخفيف

في هذا البحث أُجريت دراسة لتقييم العديد من مُحللات الجذوع العربية (**Arabic Stemmers**)، وذلك عن طريق إجراء سلسلة من المقارنات باستخدام بيانات معنونة يدوياً تُبيّن كفاءة هذه المحللات وتُشير أيضاً إلى التحسينات المحتملة لها. كما أن هذه الدراسة تقدم طريقة محسّنة للتجذير باستخدام التجذيع الخفيف (**light stemmer**) في المرحلة التحضيرية.

يُركّز هذا البحث بدايةً على فهم علم الصرف العربي الذي هو دراسة بناء الكلمات من الجذور. هذا العلم يعتمد على استخدام مجموعة من القوالب تُدعى الأوزان الصرفية، فهذه الأوزان هي سلسلة من الحروف التي تُعرِّف بدقة التغيّرات التي يمكن إجراؤها على الجذور لتوليد كلمات جديدة. يوجد نوعان من الحروف التي تُشكّل هذه الأوزان؛ الأولى هي الحروف الأصلية (**generic [original] letters**) وتمثّل حروف الجذور، مثل: ف ع ل. والثانية هي الحروف الزائدة (**augmented letters**)، وتمثّل الحروف التي يمكن زيادتها على الجذر، ويمكن جمع هذه الحروف الزائدة بكلمة «سألتمونيها».

توجد بعض الحالات التي تستوجب تعديلات إضافية وذلك حسب القواعد النحوية والحروف المتوافقة. وهذه الحالات هي؛ أولاً: النطق والتحوير (**vocalization and mutation**) وتعني تحويل الحرف من شكل إلى آخر؛ مثال ذلك تحويل الحرف الضعيف «و» في كلمة "قول"إلى الحرف الضعيف «ا» في كلمة «قال» وذلك حسب زمن (**tense**) الجملة. ثانياً: إضافة السوابق واللواحق (**Prefixes and Suffixes**)، فهذه الحروف تُعد نوعاً من الحروف

الزائدة التي يمكن إضافتها إلى بداية الكلمة أو نهايتها. حروف أخرى يمكن استعمالها مثل حرف «ك» في كلمة «كتابك» وهو غير موجود في مجموعة الحروف الزائدة. ثالثاً: الكلمات المستبعدة (Stop words)، مثل كلمة «في» التي لا تتبع قواعد الاشتقاق من الجذر- الوزن، وعموماً لها شكل ثابت. رابعاً: علامات التشكيل (Diacritics) التي تعد جزءا من الكلمة. وهي تتضمن مجموعة من الخصائص الصرفية والنحوية والصوتية المهمة. كلّ التحديات السابقة دفعت العلماء إلى وضع فرضيات مختلفة لبناء محللات الجذوع العربيّة التي تحقق التوازن بين النتائج الصحيحة وغير الصحيحة.

طُوّرت العديد من أدوات التجذير التي تستخدم مختلف الخصائص الصرفية ومنها: "Khoja Stemmer" الذي يبدأ بإزالة التشكيل وعلامات الترقيم وغير الحروف والحروف السابقة واللاحقة من الكلمة المدخلة، ومن ثم تُقارن مع مجموعة من الأوزان والجذور لحين الوصول للجذر الصحيح. وتُعد هذه الأداة الأقرب محاكاة للتجذير اليدوي غير أنها لا تشتمل على حالات خاصة مثل التشكيل. أما "Sebawi Stemmer" فيعتمد على استخدام أزواج من الجذور والكلمات لاستنتاج أوزان الكلمات والحروف السابقة واللاحقة، فمعرفة الكلمة وجذرها يُمكننا من تقسيم الكلمة إلى حروف سابقة وحروف لاحقة وجذع. بعد ذلك تتم محاذاة جذع الكلمة مع حروف الجذور لتشكيل وزن الكلمة. أما "Light10 Stemmer" فيُركّز على إزالة الحروف السابقة واللاحقة من الكلمة. وقد طُوّر هذا المحلل لإثبات فعالية التجذيع الخفيف (light stemming) في تطبيقات استرجاع المعلومات (information retrieval). أما "ISRI Stemmer" فيعمل بدايةً على تطبيع الكلمات (normalizing words) وإزالة التشكيل والحروف التي ليس لها صلة، وبعد ذلك تُزال الحروف السابقة وتُقارن مع مجموعة من الأوزان حسب طولها، وفي حال عدم وجود توافق تُزال الحروف اللاحقة وتُعاد عملية المقارنة التي تتوقف في حال كانت حروف الكلمة المتبقية تساوي ثلاثة أو أقل. "Tashaphyne Stemmer" هو محلل تجذيع خفيف يعمل على إزالة الحروف غير ذات الصلة مثل التشكيل ومن ثم يستخدم قائمتين من الحروف السابقة واللاحقة لاستخراج جذع أو جذر الكلمة.

"ElixirFM Morphological Analyzer" هو عبارة عن محلل صرفي وضيفيّ يستخدم ميزات نحوية لتمييز الدلالات المختلفة للكلمات (word secnces). إن القواعد العربيّة النحوية والصرفية مترابطة بشكل كبير، بحيث إن إضافة العديد من الحروف السابقة واللاحقة لها تفسيرات قواعدية تُساهم في عملية تشكيل الأوزان كإضافة الضمائر المتصلة. هذا المحلل الصرفي يستخدم العلاقة السابقة لتحسين عملية استخراج الجذور، إضافة إلى ذلك فإنّه يتعامل مع الحالات الخاصة مثل التحورات (mutation) باستخدم القواعد الإملائية والصوتية. يمتاز هذا المحلل الصرفي بقدرته على إيجاد جميع الجذور وربطها مع الصفات المستخرجة لتمييز الدلالات المختلفة للكلمات. "MADAMIRA Morphological Analyzer" يُقدّم هذا المحلل العديدَ من الخصائص القيّمة بما فيها استخراج جذوع الكلمات. ويتكوّن هذا المحلل من جزأين؛ الأول: MADA ويُعلّق (annotate) على الكلمة المدخلة بجميع الصفات الصرفية مثل التشكيل والتجريد (الكلمة المجردة). MADA قادر على التنبؤ بتسعة عشر ميزة صرفية باستخدام أربعة عشر آلية تصنيف للبيانات (Support Vector Machine/ SVM)، كما يتنبأ بخمسة ميزات صرفية أخرى والنموذج اللغوي المكون من عدة كلمات متجاورة (Ngram). الجزء الثاني (AMIRA) يتضمن محللا معجميا (word Tokenizer)، ووسم أقسام الكلام (POST)، وأداة تقسيم النص إلى عبارات أساسية (BPC). ويعتمد (AMIRA) على منهجية التعلم الآلي في التحليل. ويُعدّ هذا المحلل أداة مرنة توفّر العديد من الخصائص لتطبيقات أخرى مثل: الترجمة الآلية (Machine Translation) وتمييز أسماء الأشياء (Named Entity Recognition). ويُزوّدنا أيضاً بميزة التحليل الخفيف لاستخراج جذوع الكلمات، حيث يقوم بإزالة السوابق واللواحق من الكلمة. يُعدُّ هذا المحلل أداة قوية لالتقاط البيانات الكامنة بديناميكية عالية.

تُقيّم أدوات التجذيع العربيّة باستخدام اختبار (IR) القياسي، حيث أُعدّت مجموعة بيانات تتكوّن من تسع وعشرين وثيقة استُخرجت جذور كلماتها يدوياً. هذه البيانات تُعدّ جزءا من الذخيرة العالمية للغة العربيّة (International Corpus of Arabic). وتحتوي البيانات على ١٠٣٠٢ رمز، وتُعد ٨٩٤١ من هذه الرموز كلمات عربيّة، و٦٣٢٣ من هذه الكلمات لها

جذور مشتركة و٣٦٢٩ لها جذور فريدة. لكل من هذه الكلمات خصائص مرافقة لها (كالجذع والجذر) تساعد في تقييم محللات التجذيع، وهذا ما يجعل من هذه المجموعة مرجعا مثاليا في عملية التقييم.

تزوّدنا جذور الكلمات وجذوعها في اللغة العربيّة بخصائص مهمة تساعدنا في مهام الحوسبة. تُعد الدقة اللغوية (linguistic accuracy) مقياسا مهما لكفاءة المحلل الصرفي في المهام اللغوية، وتحسب الدقة اللغوية رياضياً، بحساب النسبة بين عدد جذوع الكلمات الصحيحة وعدد الكلمات المدخلة. من ناحية أخرى يمكن استخدام الجذور كوسم للكلمات، وتُجمع الكلمات المتشابهة لغوياً معاً. هناك معيار آخر لقياس الدقة اللغوية وهو حساب القدرة التصنيفية الكلية (macro) والجزئية (micro) للجذور التي يمكن حسابها عن طريق معادلات الدقة (Accuracy) والتشخيص (Precision) والاسترجاع (Recall) و(F-1).

استُخدمت جذوع مجموعة البيانات وجذورها في التجارب للتحقق من نوعين من المحللات؛ التجذيع الخفيف واستُخدم فيها (Light10 و MADAMIRA) واستخراج الجذور واستُخدم فيها (Khoja و ISRI وTashaphyne). كما أن التجارب تتحرى إمكانية الدمج بين النوعين السابقين؛ حيث استُخدم التجذيع الخفيف عمليةً مُسبقةً لاستخراج الجذور. تُبيّن التجارب أن التجذيع الخفيف باستخدام MADAMIRA حقق دقة بلغت ٩١.٧٣٪ وتحسنا مهما بنسبة ٤٤٪ أكثر من استخدام Light10. وتُبيّن التجارب أيضاً أن استخدام MADAMIRA مُعالجا مُسبقا يُحسّن من دقة استخراج الجذور ويزيد من دقة قياس التجميع (clustering) والتصنيف(classification). كما استُنتج أيضاً أنه يُمكن تحسين أداء خوارزميات التجذيع باستخدام عملية التطبيع (تبسيط النص) (normalization) ومثال ذلك: تغيير شكل حرف الهمزة (ء).

هذه الدراسة توضّح طريقة استخدام التجذيع الخفيف عمليةً تحضيريةً لتحسين مهمة استخراج الجذور. أثبتت العديد من الدراسات الأخرى أن التجذيع الخفيف له إمكانية أعلى لتحسين نتائج تطبيقات استرجاع المعلومات (IR) أكثر من استخراج الجذور. إنّ استخدام

طرق التجذيع الخفيف مع طرق استخراج الجذور من شأنه أن يبني تمثيلاً هرمياً كاملاً لكلمات اللغة العربيّة، بالإضافة إلى أن التجذيع الخفيف يُحسّن من أداء باقي أدوات التجذيع. توضّح هذه الدراسة العلاقة الطردية بين الدقة اللغوية وباقي أدوات القياس، وتوضح أيضاً نقاط القوة والضعف لكلّ من أدوات التجذيع، حيث يمكن تلافي نقاط الضعف عن طريق دمج أكثر من أداة مع بعضها.

Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland

Year of publication: 2014

MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, Ryan M. Roth

"مدامیرا" : أداة سریعة وشاملة للتحلیل الصرفي وفكّ اللبس للغة العربیّة

المقدمة

تعد عملية معالجة اللغة العربيّة عملية معقدة نتيجة للعديد من التحديات التي تواجه المختصين في هذا المجال، وتعود إلى ثلاثة أسباب رئيسة:

١- النظام الصرفي للغة العربيّة غني بالتصاريف والزوائد.

٢- اللغة العربيّة تضم الكثير من اللبس الناتج عن التشكيل الاختياري في الكتابة وشيوع الإنحراف عن المعايير الإملائية.

٣- اللغة العربيّة غنية باللهجات العامية.

في ضوء تلك التحديات تكمن الحاجة إلى أدوات تؤدي مهماتٍ أساسيةً في معالجة اللغة العربيّة الفصحى، مثل التشكيل (diacritization)، وتجريد الكلمة من الزوائد (lemmatization)، وفكّ اللبس الصرفي (morphological disambiguation)، ووسم الكلمة بأقسام الكلمات (part-of-speech tagging)، واستخراج جذع الكلمات (stemming)، وتقسيم النص العربي إلى كلمات (glossing)، فهذه المهمات غالبا ما تكون خطوات أولية أو وسطى للوصول

إلى حلّ تحدّيات أكبر تعقيدا في معالجة اللغات الطبيعية مثل تحدي الترجمة الآلية، لذلك، فإنّ أيّ أداة تقوم بتلك المهمات يجب أن تتوافر فيها السرعة والدقة وإمكانية ربطها ببرمجيات أخرى.

دراسات سابقة

هناك دراسات وجهود كبيرة عنيت بتلك المهمات للغة العربيّة الفصحى. وهناك نظام يدعى «مدى» (MADA) صُمم لمعالجة اللغة العربيّة الفصحى.

يستخدم نظام مدى محللا صرفياً لإنتاج قائمة بكلّ التحليلات الصرفية وتفسيراتها الممكنة لكلّ كلمة بحيث تغطي هذه التحليلات السمات الصرفية للكلمة مثل التشكيل، ونوع الكلمة (أقسام الكلام)، وتجريد الكلمة من الزوائد، و ثلاث عشرة سمة خاصة بتصريف الكلمة والزوائد فيها.

يطبّق «مدى» مجموعة من النماذج آليةَ تصنيف للبيانات (Support Vector Machine)، والنماذج اللغوية المتسلسلة (N grams)، لتوقع الوسوم الصرفية الأنسب لكلّ كلمة في سياقها مثل نوع الكلمة، وتجريد الكلمة من الزوائد، وجنسها، وعددها، وإسنادها (متكلم، مخاطب، غائب). يتم ذلك عن طريق ترتيب التحليلات الناتجة عن المحلل الصرفي بحساب مجموع التطابقات المتناغمة مع السمات الأنسب المتوقعة لكلّ كلمة. التحليل الصرفي الذي يحقق أعلى مجموع تطابقات لكلمة معيّنة في سياق معيّن يتم اختياره تحليلا صرفيّا متوقع لتلك الكلمة.

تضم أدوات نظام «أميرا» (AMIRA) على أداة لتقسيم النص إلى كلمات، وأداة وسم أقسام الكلمة، وأداة لتقسيم النص إلى عبارات أساسية أو ما يعرف بالمحلل النحوي السطحي (shallow syntactic parser). تعتمد تقنية «أميرا» على التعلّم الآلي المراقب (supervised learning)، ولا يوجد اعتماد واضح على المعلومات الصرفية العميقة. لذلك، وبعكس نظام مدى فإن اميرا يعتمد على بيانات سطحية ليتعلم التعميمات، وفي نسخ حديثة من اميرا أُضيف محلل صرفي ومكوّن آخر للتعرف على أسماء الأشياء (NER).

يتّخذ نظام اميرا نهجاً متعدد الخطوات في تقسيم النص إلى كلمات، ووسم الكلمة بأقسام الكلام، وتجريد الكلمة من الزوائد، بعكس نظام مدى الذي يتعامل مع كلّ تلك المهام بضربة

واحدة، ويقدّم تحليلا صرفياً أعمق من نظام اميرا، وذلك بقدرته على وسم الكلمات بطريقة توضّح علامة الإعراب والحالة الإعرابية والحالة المبنية، ولكن يترتيب على ذلك بطء سرعة البرنامج.

مداميرا

يتّبع نظام مداميرا نفس التصميم العام لنظام مدى مع بعض الإضافات المستقاه من نظام اميرا. يتم إدخال النص إلى المعالج الذي يقوم بتنظيفه وتحويله إلى هيئة (Buckwalter) المستخدم في نظام مداميرا، وبعد ذلك يُمرر النص إلى المحلل الصرفي الذي بدوره يقوم بعمل قائمة بكلّ التحليلات المحتملة لجميع الكلمات بغض النظر عن (خارج) سياقها، ثم يمرر النص والتحليلات إلى مكوِّن نمذجة السمات الذي يُطبّق نماذج آلية تصنيف للبيانات (SVM and language models) لاشتقاق توقعات مناسبة للسمات الصرفية للكلمة. تستخدم النماذج الآلية لتصنيف البيانات (SVM) لتوقّع مجموعة السمات المغلقة، أما النماذج اللغوية (language models) فتُستخدم لتوقّع مجموعة السمات المفتوحة مثل المدخلات المعجمية والتشكيل. ويحتوي النظام على عنصر خاص بترتيب قوائم التحليلات الناتجة لكلّ كلمة وإعطائها مجموعة من النقاط بناء على مدى اتفاق تلك القائمة مع توقّعات النموذج المستخدم، ثم تُصنّف التحليلات اعتمادا على تلك النقاط.

تُختار التحليلات الحائزة على أعلى نقاط لكلّ كلمة، ومن ثم تمريرها إلى مقسِّم النص ليقوم الأخير بإعداد تقسيم أو عدة تقسيمات خاصة بكلّ كلمة، ثم تُستخدم تلك التحليلات والتقسيمات من أداة تقسيم النص إلى عبارات أساسية ليقسّم النص المدخل إلى أجزاء باستخدام نموذج SVM آخر، وبنفس الطريقة تستخدم أداة للتعرف على أسماء الأشياء نموذج SVM لتمييز الأسماء وتصنيفها في النص. وعندما تنهي كلّ تلك المكوّنات عملها تُعاد النتيجة إلى المستخدم، فالمستخدم يستطيع تحديد المعلومات التي يريدها إلى جانب الكلمات والعبارات الأساسية، وأسماء الأشياء، وعلامات التشكيل، والمدخلات المعجمية، والمعاني، والسمات الصرفية، ونوع

الكلمة (اسم، أو فعل، أو حرف)، وجذع الكلمة، فجميعها تُقدّم عن طريق التحليل الصرفي الأنسب الذي يُختار من النظام.

بالإضافة إلى مضاعفة كفاءة الأنظمة السابقة، صُممم مداميرا ليكون أكثر سرعة وقابلا للتوسيع، وسهل الاستخدام والصيانة. مداميرا تدعم XML و HTTP وهذه الميزة غير متوفرة في نظامي مدى واميرا، فالنص المدخل والمخرج يمكن أن يتوفر بشكل نص عادي أو على شكل (XML).

تقسيم النص إلى كلمات

توفر مداميرا حاليا ١١ طريقة مختلفة (مخططاً) لتقسيم النص المدخل إلى كلمات. توصف هذه الطرق المختلفة بما هي العناصر التي سيتم تقسيمها أو فصلها عن الكلمة الأساسية، وما هو الشكل الذي ستُعرض فية نتائج تقسيم الكلمات. وتشير كلمة تقسيم الكلمات في النص (Tokenized) إلى أن هناك سوابق أو لواحق معينة فُصلت عن الكلمة الأساسية، كما تشير إلى أنه قد تم إجراء التعديل الإملائي اللازم للكلمة بعد فصل تلك الزوائد. أما كلمة التطبيع أو التحوير (Normalize) فتشير إلى أنه عند وجود مجموعة فرعية من الأحرف العربيّة في مقطع الكلمة فإنها تُستبدل بأحرف أو بمجموعة أحرف نموذجية، مثل حرفي الألف والياء. وأما إشارات الزوائد فهي عبارة عن رموز (غالبا ما تكون الرمز "+") تلحق بالزوائد التي فُصلت لتدل على الجهة التي تتصل منها تلك الزوائد بالكلمة الأساسية. وتوصف مخططات تقسيم النص إلى كلمات بما يُفصل من زوائد عن الكلمة الأساسية وبالصيغة الناتجة لكل كلمة أو رمز.

التقييم

لتقييم نظام مداميرا، أجري اختبار باستخدام حوالي ٢٥ ألف كلمة من اللغة العربيّة الفصحى و٢٠ ألف كلمة من اللهجة العربيّة المصرية ثم قُورنت النتائج بالنسخة الذهبية الموسومة. وقد تم التقييم اعتمادا على العديد من مقاييس الدقة (جميعها على مستوى الكلمات).

ويُذكر أنّ أحد هذه المقاييس تُقيّم نسبة الكلمات الصحيحة الناتجة من تقسيم النص باستخدام الطريقة الشائعة الاستخدام في بنك الشجيرات النحوية (ATB)، الذي يعمل على فصل جميع الزوائد المتصلة بالكلمة ما عدا (الـ) التعريف. كما يعنى هذا التقييم أيضا بنسبة الكلمات التي جُزّئت بشكل صحيح، أي بعدد مقاطع الكلمة الناتجة بغض النظر عن صحة إملائها.

تظهر مقاييس الدقة الأخرى أنّ الأنظمة السابقة لديها أداء أفضل بشكل طفيف لا يزيد عن ٠.٢٪ (مطلق) للغة العربيّة الفصحى ولا يزيد عن ٠.٦٪ للهجة المصرية. هذا الانخفاض الطفيف يعوّض عنه التحسّن الكبير في السرعة، فنظام مدامير أسرع بـ١٦-٢١ مرة من الأنظمة السابقة في معالجة اللغة العربيّة الفصحى وبـ ١٤-١٩ مرة في معالجة اللهجة المصرية. اللهجة المصرية أبطأ قليلا من اللغة العربيّة الفصحى وذلك لأنّ خطوة التحليل الصرفي لها معقدة وتنتج عنها تحليلات أكثر يتم التعامل معها من مدامير.

نتوقع تحسينات أكبر للنماذج الداخلية تكون قادرة على زيادة الدقة وتحافظ بنفس الوقت على إنتاجية معالجة الكلمات. وعند مقارنة مدامير مع نظام اميرا الأحدث فإنّ معلومات اختبار اللغة العربيّة الفصحى بالنسبة لتقسيم النص إلى كلمات كان نظام اميرا قادرا على تحقيق دقة في التقسيم تصل إلى ٩١.٤٪، وقد وصلت دقة مدامير لتجزئة الكلمة إلى ٩٩٪.

8th annual CLUK research colloquium

Year of publication: *2005*

Developing Robust Arabic Morphological Transducer Using Finite State Technology

Mohammed A. Attia

تطوير محلل صرفي عربي مرن باستخدام تكنولوجيا الحالة المنتهية

على الرغم من تميّز اللغة العربيّة في العديد من المجالات، إلا أنّها تعد من أكثر اللغات المعقدة والثرية من الناحية الصرفية (**Morphology**)، ويعد ذلك تحديا كبيراً في اللغويات الحاسوبية وتحليلها، حيث تتطلب اللغة العربيّة توافق الفعل مع الفاعل من عدة جوانب مثل العدد والنوع، وكذلك الأمر بالنسبة للعلاقة ما بين الصفة والاسم الموصوف، فيجب أن يتفقا بالعدد والجنس والحالة، هذا بالإضافة إلى حالة المثنى في اللغة العربيّة غير الموجودة غالبا في اللغات الأخرى. ولتطوير اللغة العربيّة من الناحية الصرفية كان لابد من الأخذ بعين الاعتبار بعض الاستراتيجيات التي تساعد في تحليلها وتطويرها. وتُقسم تلك الاستراتيجيات بناءً على مستوى التحليل، فمستوى التحليل الأول اهتم بتحليل سياق الكلمة واستخدام تسلسل منتظم، أما مستوى التحليل الثاني فقد اختص بتحليل الكلمات من حيث جذر الكلمة، والأنماط إضافة إلى التراكيب. ويمكن تعريف جذر الكلمة في اللغة العربيّة على أنه الشكل الأساسي للكلمة الذي يتكوّن من عدد من الحروف الساكنة، ثلاثة حروف في الأغلب، تخضع لسلسلة من التبادل مع الحروف متحركة، من انعطاف واشتقاق لتشكيل المئات من الكلمات. وبالنسبة للمستوى الثالث فقد جمع بين المستويين الأول والثاني، فقد اهتم بتحليل الكلمات من حيث جذورها، وقوالبها، وتعابيرها إضافة إلى السلاسل.

لقد اعتمد الباحث في هذه الدراسة على المستوى الأول من التحليل (**level of analysis**)، الذي يعتمد على تحليل جذوع (**roots**) الكلمات العربيّة مع الأخذ بعين الاعتبار الاختلافات

الإملائية لتلك الكلمات وفقا للقواعد المختلفة. فقد عمل الباحث على تطوير المحلل الصرفي (morphological Analyzer) على هذا الأساس. ويعد جذر الكلمة، كالشكل الأصلي للفعل، في عملية التحليل أكثر كفاءة وخصوصا في الأنظمة المتعلقة باسترجاع المعلومات، أما بالنسبة للجذع، فيُبرر استخدامه في هذا المنهج لأنه أسرع وأسهل وأكثر ملائمة بالنسبة للمحللين الذين يهدفون إلى عملية الترجمة تحديداً.

ركز الباحثون على استخدام تقنية الحالة المنتهية (finite state)، وتحمل هذه التقنية في طياتها العديد من الإيجابيات، ومن أهمها: بإمكان هذه التقنية التعامل مع القوانين المتسلسلة وغير المتسلسلة التي تُحدد كيفية ربط المحوّلات الصرفية (Transducer) مع بعضها. كما أنّ هذه التقنية تتسم بالسرعة والكفاءة، فيمكنها التعامل مع مفردات ضخمة جدا وبشكل أوتوماتيكي، فيمكن تحليل الملايين من هذه المفردات خلال ثوانٍ معدودة، بالإضافة إلى أن هذه التقنية يمكنها أن تعمل على الكثير من البرامج، فبذلك يمكن أن تخدم الكثير من التطبيقات التي تعمل عبر تلك البرامج.

يعالج النظام المقترح المقاطع الصرفية المتسلسلة والتغيرات الإملائية المتعلقة بها، حيث يجري التعامل مع هذه المقاطع الصرفية على شكل فئات مستمرة، أما بالنسبة للمتغيرات الإملائية فتُجمع متغيراتها حسب القواعد. وكأي محلل صرفي آخر، تُوفّر تغطية كافية وفعّالة للمفردات المعجمية العربيّة.

وفّر النظام كثيراً من الوقت، وهذا يرجع إلى سببين رئيسيين: الأول أن القائمين على هذا البحث استخدموا جذوع الكلمات (Stems) وليس جذورها، وثانيًا أنهم تجنبوا استخدام علامات التشكيل؛ إذ يعد إدخال علامات التشكيل عبئًا ثقيلاً على عاتق كل من مطور البرامج ومعجم المفردات. ويعالج النظام المقترح مجموعة مكونة من مئة وخمسة عشر فعلاً، وعشرة أسماء وصفات إضافة إلى مجموعة كبيرة متنوعة من حروف الجر، والأدوات والأفعال الناقصة.

تخضع الأفعال التي تُختار لبعض المواصفات، وترتبط جميع المدخلات مع بعضها ضمن قواعد معينة، فمثلا لا يمكن لأسئلة نعم أو لا (الأسئلة الاستفهامية) أن تأتي في حالة النصب

أو الأمر، كما أن ضمائر المفعول به لا يمكن أن تظهر مع الأفعال اللازمة أو مع الفعل المبني للمجهول. الأفعال المضارعة أو الماضية أو حتى أفعال الأمر ترتبط بمقاطع قبلية أو مقاطع بعدية محددة وكل منها محدد بضوابط دقيقة لا يمكن تجاوزها.

لقد تمكن النظام المقترح من تكوين ما يصل إلى ١٨٠٠ نموذج جيد للأفعال. واختُبر الفعل (شكر) الذي ولّد كمّا هائلا من الاختلافات في الشكل، ومن هذه الأشكال: يشكر، يشكران، يشكرون، تشكران، تشكرون، اشكر، اشكروا، اشكرا وغيرها. ويعد هذا الكمّ الهائل من الاختلافات في الشكل مؤشراً جيداً على ثراء العلوم الصرفية وتعقيدها في اللغة العربيّة. ويتمّ التدقيق الإملائي بشكل يدوي لضمان الجودة.

وكذلك الأمر بالنسبة للأسماء، فهناك قواعد وضوابط ترتبط بها، ومنها: ترتبط (الـ) التعريف بالأسماء لكنها لا يمكن أن ترتبط بالضمائر أو باسم غير محدد. وحروف الجر لا يمكن أن ترتبط بحالة النصب أو في جمل المبتدأ والخبر. يمكن لهذه النظام توليد ما يصل إلى ٥١٩ اسماً، مشتملاً على حالات التأنيث والتذكير إضافة إلى المفرد، الجمع والمثنى، ففي دراستهم اختُبرت بعض الكلمات مثل معلم، طالب، كتاب، كراسة وغيرها. وقد قُسّمت ضمن فئات محددة مثلا الجذع (معلم، طالب.. الخ)، المفرد المؤنث (معلمة، طالبة) المفرد المذكر (معلم، طالب)، المثنى المؤنث (معلمتان، طالبتان)، المثنى المذكر (معلمان، طالبان)، جمع المؤنث السالم (معلمات، طالبات)، جمع المذكر السالم (معلمون) جمع التكسير (طلاب)، وهكذا مع باقي الأسماء. فمن خلال اختبار تلك الكلمات والفئات تمّ الوصول إلى بعض النتائج بالنسبة للأسماء، ومنها: الأسماء جميعها تحتوي على مثنى (مذكر أو مؤنث أو كلاهما)، بعض الأسماء يمكن أن تحتوي على ثلاثة أنواع من الجمع (مؤنث سالم، مذكر سالم وجمع تكسير)، علما بأن جمع التكسير يُدخل بشكل منفصل من اللغويين. كما أنّ هناك بعض الأسماء التي يختلف فيها مذكرها عن مؤنثها تماما مثل كلمة (ثور) وهي مذكر ومؤنثها (بقرة). ووجود جمع التكسير أحد التحديات التي واجهت الدارسين لهذا النظام، إذ كان يتوجب عليهم تحديد تلك الجموع وإدخالها حتى يتقبلها النظام المقترح.

ومن ميزات هذا النظام أنه يمكنه دمج أكثر من عنصر، ومن أهم هذه العناصر:

١- أداة الترميز: وتقوم على إعطاء رمز لكل سطر ومعالجة الكلمات التي تحتوي على أكثر من معنى.

٢- المحلل الصرفي.

٣- المخمّن: ويعمل على مهمتين أساسيتين، فتقوم مهمته الأولى بالمحافظة على سير المحلل لإعطاء التحليلات وعدم فشلها، أما المهمة الثانية فتكمن في إضافة مصطلحات جديدة إلى المعجم الأساسي.

٤- منظّم علامات التشكيل: على الرغم من أن النظام لا يتعامل غالبا مع النصوص المشكلة (التي تحتوي على حركات التشكيل)، إلا أن هذا العنصر وُجد في حال وجود نصوص مشتملة على حركات، والهدف من استخدامها هو منع النظام من الفشل في تقديم تحليل صحيح.

٥- المحلل الإملائي: وهذا العنصر يتعامل مع الاختلافات الإملائية الشائعة والمتنوعة إضافة إلى الأخطاء الإملائية. ومن أهم تلك الاختلافات أو الأخطاء ما يتعلق بكتابة الكلمات التي تحتوي على أحد الحروف مثل الألف المقصورة والياء(ى و ي)، الألف الممدودة والألف المقصورة(ا و ى)، الحروف المحتوية على همزات(ئ، ء، أ، ؤ)، حروف الألف بأشكالها المختلفة (أ، ا، إ)، الهاء المربوطة والتاء المربوطة (ه، ة) وغيرها من الحروف التي يمكن أن يكون فيها لبس أو غموض في كتابتها.

1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), Sharjah, 2013, pp. 1-6. doi: 10.1109/ ICCSPA.2013.6487311

***Year of publication:** 2013*

SALMA: Standard Arabic Language Morphological Analysis

Majdi Sawalha, Eric Atwell, Mohammad A. M. Abushariah

المحلل الصرفي المعياري للغة العربيّة: "سلمى"

الملخص

تُعد المحللات الصرفية معالجاتٍ للنصوص تحتاجها بعض التطبيقات المتخصصة لإنجاز مهامها، أحدها **SALMA** وهو مجموعة من الموارد والأدوات المفتوحة المصدر لتوسيع مجال تحليل بنية الكلمة العربية؛ تحديداً التحليل الصرفي، وذلك لتحليل نصوص ذخائر اللغة العربيّة. و**SALMA-Tagger** هو محلل صرفي تفصيلي يعتمد المعلومات اللغوية المستخرجة من كتب قواعد اللغة العربيّة التقليدية، ويعتمد أيضاً على معلومات سابقة من مصدر معجمي واسع التغطية سمي (**SALMA-ABCLexicon**).

المقدمة

يعدّ التحليل الصرفي خطوة سابقة للعديد من تطبيقات تحليل النصوص، التي اجتذبت الباحثين من تخصصات مختلفة كاللغويات (اللغويات الحاسوبية ولغويات الذخائر)، والذكاء الاصطناعي، ومعالجة اللغات الطبيعية، لتحليل النصوص صرفياً ونحوياً للغات عديدة ومنها اللغة العربية. واعتمد باحثون كثر مناهج مختلفة في التحليل النحوي والصرفي للعربية، فهناك العديد من الأنظمة التي تتباين في تعقيدها، كالتجذيع الخفيف (**light stemmer**)،

وأنظمة استخراج الجذور (root extraction systems)، وأنظمة تجريد الكلمة من الزوائد (lemmatizers)، والمحللات الصرفية المعقدة (complex morphological analyzers)، وبرامج وسم أقسام الكلام (part-of-speech taggers)، والمحللات النحوية (parsers).

حوسبة الصرف العربي

العربيّة لغة حية تنتمي إلى اللغات السامية، ومن أهم خصائصها أنّ الصرف فيها غير إلصاقي (nonconcatinaive morphology) حيث تشتق الكلمات من الجذور. ودرست النظريات اللغوية الحديثة الصرف العربي وعملية اشتقاق الكلمات من وجهتي نظر: على أساس الجذر، أو على أساس الجذع. ويتضمن التحليل الصرفي للعربيّة التطبيقات الحاسوبية التي تحلل بنية الكلمات العربيّة، وتنتج جميع التحليلات الممكنة للكلمة المكتوبة، وتجرى هذه العمليات إما على أساس الشكل أو على أساس الوظيفة، والمحلل الصرفي يحتاج إلى تطوير كلا النموذجين. وتتضمن هذه العمليات، تقسيم النص إلى كلمات (tokenization)، والتدقيق الإملائي (spell-checking)، واستخراج جذع الكلمات (stemming)، وتجريد الكلمة من الزوائد (Lemmatization)، ومطابقة الهيئة (pattern matching)، والتشكيل الآلي (diacritization)، والتنبؤ بالخصائص الصرفية للوحدات الصرفية للكلمات (predicting the morphological features of the word's morphemes)، ووسم أقسام الكلام (part-of-speech tagging)، والتحليل النحوي (parsing).

واستخدمت المحللات الصرفية عدداً من الطرق نذكر منها التحليل الصرفي المعتمد على الوحدات اللفظية للكلمات (Syllable-Based Morphology (SBM)) الذي يعتمد تحليل الوحدات اللفظية للكلمة. والطريقة التي تعتمد على وزن (Root-Pattern Methodology) الكلمة وجذرها في التحليل. والتحليل الصرفي الذي يعتمد على جذع الكلمة (Lexeme-based Morphology) حيث يشكل جذع الكلمة المستخرج معلومة مهمة. وتعتمد الطريقة الأخيرة على مصدر معجمي يحتوي جذع الكلمة (Stem-based)، وعلى قواعد اشتقاق الكلمات.

إذاً، كيف نقوم بتطوير التحليل الصرفي وإيجاد الجذع؟ بُني واستُخدم مصدرمعجمي واسع التغطية (SALMA- ABCLexicon) لزيادة دقة المحللات الصرفية، ولبناء هذا المصدر المعجمي، تم اختيار ٢٣ معجماً متوفرة إلكترونياً. وهناك ثلاثة عوامل حددت توجهنا لبناء المصدر المعجمي واسع التغطية، هي: غياب محتوى عربي كبير ومفتوح المصدر، وغياب برنامج توليد مفتوح المصدر، ووجود مشاكل برمجية في برنامج التوليد الصرفي.

معايير التحليل الصرفي والنحوي (Morphosyntactic):

التجربة الأولية للمحللات الصرفية وبرامج استخراج الجذع أو الجذر أشارت إلى عدم وجود معايير وإرشادات لعمليات وسم النصوص العربيّة صرفياً ونحوياً، وتُعد هذه المعايير والإرشادات متطلباً سابقاً لعمليات الوسم للذخائر اللغوية.

في اللغات الغنية صرفياً مثل العربيّة، يجب أن تعرّف مجموعة أقسام الكلام مسبقاً بالاعتماد على الخصائص الصرفية التي تميز بنية الكلمة، وعليه فقد طُوِّرت مجموعة وسم أقسام الكلام (SALMA-Tag Set)، التي تمتاز بما يأتي:

١- تتضمن مجموعة وسم أقسام الكلام الخصائص الصرفية للغة العربيّة.

٢- يوضّح الشرح التفصيلي لمجموعة وسم أقسام الكلام SALMA وصفاً تفصيليا لكل خاصية صرفية ويحدد القيم الممكنة لكل خاصية صرفية.

٣- يحتوي الوسم على ٢٢ رمزاً، كلّ موقع يمثل خاصية صرفية، وكل رمز في موقع معين يمثل قيمة لهذه الخاصية الصرفية، ويعني الرمز (-) في موقع معين أن الخاصية الصرفية في ذلك الموقع لا تنطبق على الكلمة المحللة.

٤- لا تخضع مجموعة وسم أقسام الكلام SALMA إلى خوارزمية أو نظرية محددة أو طريقة وسم محددة، ويمكن ربط مجموعات الوسوم الأخرى إلى هذا المعيار، وعليه تكون الدراسات المقارنة للمحللات الصرفية والدخائر الموسومة أسهل.

تقيّم مجموعة وسم أقسام الكلام (SALMA) من خلال طريقتين، الأولى تكون من خلال

اقتراحها كمجموعة معيارية لمجتمع حوسبة اللغة العربيّة، واستُخدمت من بعض أنظمة معالجة اللغة العربيّة:

ا- ستُخدامت في المحلل الصرفي (SALMA) لترميز الخصائص الصرفية لكل وحدة صرفية في الكلمة.

- استخدمت جزئياً من مجموعة وسم أقسام الكلام في المحلل الصرفي وموسم أقسام الكلام "قطوف".
- إقرارها معيارا لتقييم المحللات الصرفية العربيّة وبناء معيار ذهبي لتقييم المحللات الصرفية وموسمات أقسام الكلام للنص العربي.
- أما الطريقة الثانية، فقد طبقت منهجية تجريبية لتقييم مجموعة وسم أقسام الكلام (SALMA) التي أظهرت أنه يمكن تطبيقها لوسم الذخيرة العربيّة.

التطبيق والتنفيذ

المحلل الصرفي والنحوي تطبيق مهم وأساسي في معالجة اللغة الطبيعية، حيث يمكن دمجها مع تطبيقات مهمة واسعة من تطبيقات معالجة اللغات الطبيعية. يعد موسم SALMA محللا صرفيا تفصيليا مفتوح المصدر يدمج الموارد والمعايير المطورة في هذا البحث معاً. كما يعتمد أيضاً على القوائم المدخلة مسبقاً مثل قوائم الجذور، والسوابق واللواحق، والأوزان الصرفية، والكلمات الوظيفية، وجموع التكسير، وقوائم أسماء الأشياء .. الخ، التي استُخرجت من كتب قواعد اللغة العربيّة التقليدية. وقد طُوّر هذا المحلل الصرفي لتحليل الكلمات وتحديد خصائصها الصرفية، ويستخدم هذا المحلل نموذجاً لتجزئة الكلمة المحللة إلى خمسة أجزاء، كما عُرّفت مجموعة وسم أقسام الكلام، حيث يُعيّن وسم صرفي مكوّن من ٢٢ خاصية صرفية لكل وحدة صرفية في الكلمة. يتكوّن المحلل الصرفي (SALMA) من مجموعة من الوحدات التي يمكن استخدامها بشكل مستقل لتنفيذ مهام، كاستخراج الجذور والجذوع، وتجريد الكلمات، ومطابقة أوزان الكلمات. ويمكن استخدام هذه الوحدات معاً لإنتاج التحليل الصرفي التفصيلي للكلمات.

التقييم

أشار تقييم المحلل الصرفي **SALMA** إلى أن طرق التقييم للمحللات الصرفية لم توحَّد بعد، لذا طُوّرت معايير متفق عليها لتقييم المحللات الصرفية للنص العربي يعتمد على خبرتنا العملية والمشاركة في المؤتمرات والمسابقات. كما طُوّر معيار ذهبي (**SALMA-Gold Standard**) لتقييم المحللات الصرفية للغة العربيّة، مكوّن من ١٠٠٠ كلمة من نصوص القرآن الكريم و١٠٠٠ كلمة من نصوص الذخيرة العربيّة المعاصرة.

قُيّم المحلل الصرفي (**SALMA**) بمقارنة نتائج التحليل بما يقابلها في المعيار الذهبي، وركّز التقييم على قياس دقة التحليل للخصائص الصرفية لكلّ وحدة صرفية (**Morpheme**) في الكلمة، فأظهرت النتائج ٥٣.٥٪ من نصوص القرآن الكريم و٧١.٢١٪ من النصوص الأخرى حُللت بشكل صحيح، وقد دُققت باستخدام تقنية المقابلة الدقيقة والكاملة للوسم الصرفي.

هذه النتائج بشكل عام أظهرت أنه بالإمكان تطبيق تحليل صرفي تفصيلي لنصوص اللغة العربيّة، كما أظهرت النتائج قدرة المحلل الصرفي (**SALMA**) على تحليل نصوص من أنواع مختلفة، وتحليل النصوص المشكولة وغير المشكولة، كما أنه بالإمكان استخدام المحلل الصرفي (**SALMA**) لوسم الذخيرة العربيّة بأقسام الكلام والتحليل الصرفي التفصيلي لجميع كلمات الذخيرة العربيّة. كما أظهرت النتائج أنه بالإمكان تطوير محلل صرفي لنصوص اللغة العربيّة بالاعتماد على نظام معرفي معتمد على نظام لغوي لتحديد قيم الخصائص الصرفية التفصيلية للكلمات المحللة.

أما نتائج تقييم الخصائص الصرفية المنفردة، فهي مفيدة لمستخدمي نظام **SALMA** لمعرفة نسبة الدقة لكل خاصية صرفية، حيث كانت نسبة الدقة عالية ل ١٥ خاصية صرفية، وتراوحت دقة التحليل من ٩٨.٥٣٪ - ١٠٠٪ لنصوص الذخيرة العربيّة المعاصرة، وبين ٩٠.١١٪-١٠٠٪ لنصوص القرآن الكريم. والتصنيفات السبع المتبقية حققت نسبة دقة أقل: من ٨١.٣٥٪-٩٧.٥١٪ لنصوص الذخيرة العربيّة المعاصرة و٧٤.٢٥٪-٨٩.٠٣٪ للنص القرآني.

الخلاصة

طوّر عدد من الحاسوبيين واللغويين خوارزميات لحصر المشاكل في الوسم الصرفي الآلي للنص العربي فعالجت هذه الدراسة المحللات الصرفية الحالية، وعملت التجارب على اكتشاف التحديات والنظريات العملية. الجزء العملي ضم تطوير موارد لتحسين الدقة لهذه الأنظمة، فهذه الموارد يمكن إعادة استعمالها في تطبيقات لغوية تعالج النصوص العربيّة كما تضم تطوير معايير للتحليل الصرفي للغة العربيّة معتمداً على معرفة لغويةٍ مستنبطة من قواعد اللغة العربيّة التقليدية الراسخة.

وفي النهاية، الموارد والمعايير جُلبت معاً لتطوير نظام **SALMA** المحلل الصرفي التفصيلي لنصوص اللغة العربيّة من مختلف الأنواع والأشكال.

IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 2, No 1, March 2014

Year of publication: *2014*

Arabic Roots Extraction Using Morphological Analysis

Aymen Abu-Errub, Ashraf Odeh, Qusai Shambour, Osama Al-Haj Hassan

استخراج الجذور العربيّة باستخدام التحليل الصرفي

تتميز اللغة العربيّة بتركيبها المعقّد على أساس نمطها المعتمد على الجذور (**Roots**). ويعدُّ استخراج الجذر أحد أهم المواضيع في تطبيقات معالجة اللغات الطبيعية مثل استرجاع المعلومات، ومعالجة النصوص، والترجمة الآلية، ووسم الكلام، إلخ. تقدم هذه الورقة طريقة لاستخراج الجذور الثلاثية للكلمات العربيّة، التي تعمل مـن جذور ذات ثلاثة أحرف ساكنة، من خلال إزالة السوابق (**prefixes**) واللواحق (**suffixes**) واستخدام قائمة من الأوزان الصرفية (**morphological weights**). تظهر النتائج التجريبية المستندة إلى قائمة تضم١١ حالة للجذور فعالية الطريقة المقترحة بمعدل نجاح قدره ٩٤٪.

تعتمد الطريقة المقترحة لاستخلاص جذور الكلمات على خطوات أساسية تبدأ باستخدام الخوارزمية الجذرية لإزالة اللواحق والسوابق التي تضاف إلى الجذر، ثم تُحذف اللواحق من خلال المقارنات مع التحليل الصرفي مع الأخذ بعين الاعتبار عدم استخدام أيّ جذور خاصة للمقارنة.

الخطوة الأولى في الطريقة المقترحة هي تجذيع الكلمات (Stemming)

تهدف هذه العملية إلى حذف سوابق (**prefixes**) ولواحق (**suffixes**) الكلمات. تبدأ العملية عن طريق حذف الحروف السابقة من الكلمة، وبعد الحذف، يُحسب عدد أحرف الكلمة للتأكد من أنه ما زال هناك أكثر من ثلاثة أحرف. إذا كان عدد الأحرف بعد الحذف

أقلّ من ثلاثة، تُلغى عملية الحذف. كذلك الأمر تُكرر عملية حذف أحرف السوابق هنا لحذف أحرف اللواحق. وبعد الانتهاء من الحذف، يُحسب عدد أحرف الكلمة للتأكد من أنها لا تزال أكثر من ثلاثة أحرف. إذا كان عدد الأحرف بعد الحذف أقل من ثلاثة، تُلغى عملية الحذف.

الخطوة الثانية: خوارزمية استخراج جذور الكلمات (root extraction)

لاستخراج الجذر، يجب مقارنة الكلمة الناتجة من عملية التجذيع بوزنها الصرفي المناسب، ثم تُحذف الحروف الإضافية من الكلمة، وتقابل الأحرف الإضافية في الوزن الصرفي. وتتكون خوارزمية استخراج الجذور من عدة خطوات هي:

١- تعريف الرموز المستخدمة في الخوارزمية.

٢- قراءة عدد حروف الكلمة وتخزينها في مصفوفة.

٣- قراءة طول أول عنصر من الوزن الصرفي وتخزينه، فإذا كان طول الكلمة الأصلية مساويا لطول الوزن الصرفي: قم بمقارنة الحروف الزائدة بين الوزن الصرفي والكلمة نفسها.

٤- اذا كانت الحروف الزائدة في الكلمة الأصلية مطابقة لحروف الوزن الصرفي الزائدة، تقوم الخوارزمية بحذف الحروف الزائدة من الكلمة واعتبار ما تبقى هو جذر الكلمة.

٥- إذا لم يحدث تطابق في الخطوة ٤، تُكرر الخطوة ٣ مرة أخرى لحين الوصول إلى مطابقة بين أحد الأوزان الصرفية والكلمة الأصلية.

Word Structure, Volume 6 Issue1, Page43-99, ISSN 17501245

Year of publication: *2013*

A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging

Majdi Sawalha, Eric Atwell

مجموعة الوسم القياسية التي تفسر الخصائص الصرفية التقليدية للغة العربيّة لوسم أقسام الكلام

تتضمن مجموعة الوسوم الصرفية (SALMA) الخصائص الصرفية الراسخة في اللغة العربيّة والقواعد التقليدية، على شكل مجموعة رموز دقيقة وواضحة. في هذا البحث عُرضت المعايير القياسية لمجموعات وسم أقسام الكلام للغة الإنجليزية وبعض اللغات الأوروبية، ثم عَرض البحث (١) دراسة مسحية للذخائر العربيّة، (٢) وبرامج وسم أقسام الكلام للنصوص اللغة العربيّة، (٣) والتقاليد الراسخة للتحليل الصرفي في اللغة العربيّة. ودرس البحث مقارنة مجموعات وسم أقسام الكلام للغة العربيّة، ومراجعة المعايير القياسية العامة لتصميم مجموعة وسوم لوسم الذخائر اللغوية. في لغةٍ غنيةٍ صرفياً كاللغة العربية، يجب أن تُعرَّف مجموعة وسم أقسام الكلام بالخصائص الصرفية التي تصف بنية الكلمة. وفي هذا البحث وُصفت مجموعة وسم الخصائص الصرفية (SALMA) بشكل مفصل، بتفسير جميع الخصائص الصرفية وقيمها وتوضيحها. وفي هذا التحليل، يتكوّن الوسم الواحد من ٢٢ رمزا، يمثل كلّ موقع أحد الخصائص الصرفية، ويمثل كلّ رمز في موقع معين قيمة أو صفة هذه الخاصية الصرفية. ويعبر الرمز (-) عن عدم تطابق الخاصية الصرفية للكلمة المحللة. ويمثّل الموقع الأول في الوسم الصرفي أقسام الكلام الرئيسية، وهي:

الاسم، والفعل، والحرف، والإضافات، وعلامات الترقيم. والأخيرتان هما امتداد لأقسام الكلام التقليدية لما تتطلبه النصوص العربيّة الحديثة.

الرمز في الموقع الثاني، والثالث، والرابع يمثّل أقسام الكلام الفرعية، وتميز قواعد اللغة العربيّة بين ٣٤ نوعاً من الأسماء، وثلاثة أنواع من الأفعال، و٢١ نوعاً من الحروف أو الأدوات. تمثّل المجموعة التالية من الرموز الخصائص الصرفية: الجنس في الموقع (٧)، العدد في الموقع (٨)، الإسناد في الموقع (٩)، الصرف في الموقع (١٠)، الحالة الإعرابية للاسم أو الفعل في الموقع (١١)، علامة الإعراب أو البناء في الموقع (١٢)، المعرفة والنكرة في الموقع (١٣)، المبني للمعلوم والمبني للمجهول في الموقع (١٤)، المؤكد وغير المؤكد في الموقع (١٥)، اللازم والمتعدي في الموقع (١٦)، العاقل وغير العاقل في الموقع (١٧)، التصريف في الموقع (١٨). وجاءت آخراً أربعة رموز تمثّل معلومات صرفية تفيد معالجة النصوص العربيّة، (على الرغم من عدم اعتبار هذه الخصائص ضمن الخصائص الصرفية التقليدية): المجرد والمزيد في الموقع (١٩)، عدد أحرف الجذر في الموقع (٢٠)، بنية الفعل في الموقع (٢١)، وأقسام الاسم تبعا للفظ آخره في الموقع (٢٢). صُممت مجموعة وسم أقسام الكلام (SALMA) لتكون عامة وغير مرتبطة أو مصممة لخوارزمية أو نظرية محددة، ومن الممكن ربط مجموعات وسم أقسام الكلام الأخرى بهذه المجموعة المعيارية، فمن الممكن تسهيل المقارنة بينها، وإعادة استخدام برامج وسم أقسام الكلام، والذخائر الموسومة.

Journal of Computational Linguistics Research Volume 2 Number 1 March 2011

Year of publication: *2011*

A Computational Feature-Based Morphological Analysis and Generation of Modern Standard Arabic

Mourad Gridach, Noureddine Chenfour

التحليل الحاسوبي الصرفي المستند للمزايا وتوليد لغة عربيّة حديثة

اكتسب التحليل الصرفي العربي تركيزاً محورياً ولفترة طويلة من أبحاث معالجة اللغات الطبيعية، من أجل تحقيق الفهم الآلي للغة العربيّة. ويعدّ التحليل الصرفي العربي أداة مهمة في جميع مجالات البحث العلمي والصناعة التي تتطلب معرفة الهيكل الداخلي للكلمات العربيّة. فالتحليل الصرفي العربي بجودة عالية الذي يمثل التحدي الأول هو المرحلة الأساسية في العديد من تطبيقات معالجة اللغات الطبيعية، مثل استرجاع المعلومات والترجمة. والتحدي الثاني يهتم باستخدام التحليل الصرفي أنظمة الترجمة الآلية، فقد أثبتت الدراسات السابقة أن أنظمة الترجمة الآلية التي تعتمد على التحليل الصرفي التفصيلي تكون ذات أداء أفضل. وتزيد أهمية التحليل الصرفي في أنظمة الترجمة من وإلى لغات غنية صرفياً كاللغة العربيّة. والتحدي الثالث هو اعتبار التحليل الصرفي الخطوة الأولى للتحليل النحوي.

تعرض هذه الدراسة طريقة آلية للتحليل الصرفي تعتمد على الخصائص الصرفية للغة العربيّة بعدة مستويات عدّة: الأسماء والأفعال تتميز بطريقة تمثيها على شكل مصفوفة من الجذور والأوزان. تشتق الأسماء والأفعال العربيّة من الجذور والأوزان التي تنتج الكلمة المجرة، ومن ثم تُضاف السوابق واللواحق لإنتاج مفردات صحيحة.

أحد الأعمدة المهمة في التحليل الصرفي هو المعجم الذي يحوي مجموعة من الأشكال المعجمية (Lexical forms) الصحيحة (المفردات)، وهو الخطوة الأولى نحو بناء محلل صرفي قوي يكون هو الواجهة الأمامية لكثير من أنظمة معالجة اللغات الطبيعية. وهناك نوعان

من الجوانب التي تسهم في تحسين نوعية المعجم: الجانب الأول هو عدد مدخلات المعجم، والجانب الثاني يتعلق بالثراء في المعلومات اللغوية الواردة فيه. وفي هذه الأيام، تُطبّق طريقة جديدة للتمثيل والتصميم والتطبيق لمعاجم تعتمد إطار الترميز المعجمي (**Lexical Markup Framework** (**LMF**)) وهو الإطار المعياري (**ISO**-٢٤٦١٣) لمعالجة اللغات الطبيعية والمعاجم. أما طريقتنا في التمثيل المعجمي تعتمد (**XMODEL**) وهي لغة التعريف الصرفي المعتمدة على لغة التوصيف الموسعة (**XML-based Morphological Definition Language**).

المنهجية المقترحة في هذا البحث تعتمد أتمتة الصرف العربي، حيث تعد هذه الطريقة من بين أكفأ الطرق للتحليل الصرفي، وتستعمل هذه الآلة للتوليد والتحليل الصرفي، وتعدّ الكلمة مقبولة من الآلة الصرفية إذا كانت الكلمة تنتمي إلى كلمة عربيّة صحيحة. ويحتاج تطبيق الآلة الصرفية إلى وجود المعجم. وقد استُخرجت القواعد الصرفية من المعجم ومن ثم طُبّقت على شكل آلة صرفية لكل قاعدة.

قُيّم المحلل الصرفي بمقارنته مع أفضل المحللات الصرفية للغة العربيّة (**ElixirFM**) و(**Xerox**). أظهرت عملية تقييم العمل أن المحلل الصرفي كان قوياً فيما يتعلق بالخصائص الصرفية التي يقدمها، وتجعل النظام مفيداً لمعظم تطبيقات معالجة اللغات الطبيعية، على عكس المحاولات الأخرى التي كانت تقصد تطبيقات محددة. بالإضافة إلى ذلك، المحلل الصرفي المقدم يعطي المزيد من المعلومات الإضافية حول كل كلمة من ناحية تحليلها بدقة. تضمّن التقييم أيضاً معالجة الكلمات في مجموعة مختارة من النصوص، بحيث كانت المدخلات تحتوي على صيغ مختلفة من الكلام (الأفعال والأسماء والحروف)، ثم حُسب نجاح كلّ محلل من ناحية عدد الكلمات مع عدد الحلول / عدد الكلمات.

وبما أنّ اللغة العربيّة لغة عالية التصريف والاشتقاق، فستكون هناك دائما إمكانيات لتحسين منهجيات المعالجة الحاسوبية وأدواتها.

In: Soudi A., Bosch A., Neumann G. (eds) Arabic Computational Morphology. Text, Speech and Language Technology, vol 38. Springer, Dordrecht

Year of publication: *2007*

Light Stemming for Arabic Information Retrieval

Leah S. Larkey, Lisa Ballesteros, Margaret E. Connell

استخدام التجذيع الخفيف لاسترجاع المعلومات في النصوص العربيّة

تدرس هذه الورقة البحثية أثر تطبيق إحدى طرق المعالجة القبليّة (**preprocessing**) المطبقة على النصوص العربيّة وهي الجذيع الخفيف ((**Light Stemming** على دقّة استرجاع المعلومات، وفي هذا السياق نذكر أن التجذيع هو عبارة عن إزالة زوائد الكلمات وإرجاعها إلى الجذر، مثلا كلمة (مهندسون) يمكن تقليم زوائدها وإرجاعها إلى الأصل (هندس) أو (مهندس)، وذلك اعتمادا على طريقة التجذيع المتبعة.

تعد قضية التشكيل الصرفي مشكلة حقيقية لمعالجة اللغة العربيّة لأنها شديدة التأثّر بالحركات. ومع ذلك، وجدنا أن الحل الكامل لهذه المشكلة غير مجدٍ لاستعادة المعلومات بفعالية. لذلك، فإن التجذيع الخفيف ((**Light stemming** يسمح باسترجاع معلومات جيدة بشكل ملحوظ من دون تقديم تحاليل مورفولوجية معقدة. وقام الباحثون بتطوير العديد من المجذّعات الخفيفة للغة العربيّة وقُيّمت فعاليتها لاسترجاع المعلومات باستخدام بيانات **TREC** العامة، وعند المقارنة بين المجذعات الخفيفة المستخدمة، أظهر المجذع الخفيف ١٠ تفوقاً واضحاً على الطرق الأخرى. وقد أُدرج في مجموعة أدوات (**Lemur**) ليمور وأصبح يستخدم على نطاق واسع لاسترجاع المعلومات العربيّة.

ركزت الدراسة في نهجها على المقارنة بين أنواع التجذيع الخفيف المختلفة وبين المحللات الصرفية، حيث قارنت بين استخدام محلل بوكوولتر (**Buckwalter**) الصرفي، والمحلل المعجمي (**tokenizer**) دياب (**Diab**) وبرنامج وسم أقسام الكلام (**part of speech tagger**).

محلل **Buckwalter** يأخذ الكلمات العربيّة مع أحرف العلة القصيرة أو دونها، وينفذ التحليل المورفولوجي والوسم **POS** باستخدام ثلاثة قواميس وثلاثة جداول توافقية. تسرد القواميس الثلاثة البادئات المحتملات والجذور العربيّـة (**stems**) واللواحق (**suffixes**) المحتملة. أما محلل دياب فهو أداة متوفرة على شبكة الإنترنت وقد اشتق منها أربعة إصدارات وهي:

١. دياب التحليل الصرفي والمعجمي بهدف إزالة المقاطع المغلقة.

٢. تطبيق محلل دياب مرتين لإزالة اللواحق في التجذيع الخفيف ١٠.

٣. تطبيق دياب بعد الخطوة السابقة لإزالة اللواحق الخاصة بالجموع.

٤. تطبيق دياب على الخطوة الثالثة بهدف إزالة علامات الاستفهام أو الصفات وغيرها.

وقد أثبتت الدراسة أن للتجذيع تأثيراً كبيراً على استرجاع المعلومات العربيّة، أكبر بكثير من التأثير الموجود لمعظم اللغات الأخرى، حيث قدمت الدراسة تحسينات وصلت إلى ١٠٠٪ في متوسط الدقة بسبب العمليات الجذرية والعمليات ذات الصلة، وتأثيرًا أكبر لاسترجاع اللغة من القاموس. هذا التأثير الجذري كان كبيراً، مقارنةً بالعديد من الدراسات الجذرية الأخرى. وقد أجابت الدراسة المقدمة عن الاستفسار التالي: لماذا قدم المقلّم الخفيف نتائج أفضل من أيّ محلل صرفي متكامل؟ وكانت الإجابات على النحو الآتي:

أولاً، يرتكب المحللون الصرفيون أخطاء، لا سيما على الأسماء.

ثانياً، قد لا تتمكن نماذج الاسترجاع الحالية من استخدام المعلومات المقدمة عن طريق التحليل الصرفي.

ثالثًا، التجذيع الخفيف يعد طريقة رصينة لأنه لا يتطلب جملا كاملة لأنه يتعامل مع الكلمات، يكفي لاسترجاع المعلومات أن تُخلط أكثر أشكال الكلمات تكرارًا.

Computer Science, 14(2), 327

***Year of publication:** 2013*

Towards a new Approach for Arabic root extraction: Exploit relations between the word letters and their placement in the word for Arabic root extraction

Fatma Abu Hawas

نحو منهجية جديدة لاستخراج الجذور العربيّة: استخدام العلاقات بين حروف الكلمة ومواقعها فيها لاستخراج الجذر العربي

عرضت هذه الدراسة منهجاً جديدا لاستخراج جذر وحيد للكلمة العربيّة دون الاعتماد على قواعد بيانات للجذور والأوزان أو الرجوع لقوائم اللواصق للكلمات العربيّة. وبعكس خوارزميات التجذيع التي تعتمد على القواعد، تحاول هذه المنهجية توقع مواقع أحرف الجذر في الكلمة واحداً تلو الآخر معتمدة على القواعد والعلاقات بين أحرف الكلمة ومواقعها فيها. تركز هذه الدراسة على جزأين في المنهجية. الأول يقدم بعض القواعد للتمييز بين أداة التعريف (الـ) والمقطع (الـ) الذي يمكن وجوده في أي كلمة عربيّة. والثاني هو تصنيف الحروف العربيّة إلى مجموعات حسب مواقعها من الكلمة.

وكان تقييم المنهج باستخدامه في تحليل كلمات القرآن الكريم، وأظهرت نتائج التقييم أن هذه الخوارزمية واعدة.

المقدمة

تحظى معالجة اللغات الطبيعية حاسوبيا باهتمام كبير، وقد وضع الباحثون تطبيقات كثيرة في هذا المجال. ويعد استخراج الجذور والجذوع من أهم التطبيقات في معالجة اللغات الطبيعية، لأنه عنصر أساسي في نجاح تطبيقات لغوية عديدة. إلا أنّ عملية استخراج جذوع الكلمات العربيّة باستخدام قواعد لغوية صرفية ما زالت مهمة صعبة.

وتحاول هذه الورقة الإجابة عن سؤال الباحثين التالي:

- هل يمكننا استخراج جذر الكلمة العربيّة حاسوبيا باستخدام العلاقات بين حروف الكلمة ومواقعها في الكلمة؟

وصف المنهج

يمكن وصف هذا المنهج بأنه أعمى [غير مسترشد] لمحاولته إيجاد جذر الكلمة دون الرجوع إلى قائمة معدة سلفا للجذور العربيّة، أو قائمة للكلمات وأوزانها، أو قائمة بالسوابق واللواحق. ويقوم المنهج على تصنيف الحروف العربيّة استناداً إلى مواقعها من الكلمة، وإعطاء كل حرف قيمة ٠ أو ١ لتحديد إن كان من حروف الجذر أم لا.

تقسيم الكلمات والحروف

تقسم كلّ كلمة إلى ثلاثة أقسام (S3،S2،S1)، طول المجموعة الأولى S1 والمجموعة الثالثة S3 يحسب من قسمة طول الكلمة على ٣ مقرباً إلى أقرب عدد صحيح. وطول الجزء الثاني S2 يحسب بما تبقى من الكلمة من حروف. وتصنف حروف الكلمات العربيّة إلى ثماني مجموعات. وأثناء تنفيذ الخوارزمية يعطى كلّ حرف في الكلمة قيمة (٠، -١، ١) لتمثيل حالة وجوده في جذر الكلمة.

التعامل مع الكلمات التي تحتوي على (ال) و الكلمات التي تحتوي على علامة الإدغام (لل)

أداة التعريف (ال) تتصل بالاسم وتأتي سابقةً، وبعكس بقية الأدوات لا تأتي منفصلة أو مستقلة، وإنما تأتي مؤشراً للتعريف. ويمكن التمييز بينها وبين المكون الأساسي للكلمة من خلال فحص الحالات الآتية:

١- تُعد ال جزءا من الكلمة عندما يتبعها حرف شمسي غير مضعف.

٢- إذا سبق ال حرف واحد، وكان الحرف الذي يسبق ال أحد الحروف (ك، و، ف، ب) فهي جزء أساسي من الكلمة على ألا يتجاوز طول الكلمة خمسة حروف.

٣- إذا سبق ال حرفان كلاهما ينتميان لمجموعة الحروف (ك، ف، و، ب) فهي جزء أساسي من الكلمة.

٤- إذا سبق ال أكثر من ثلاثة حروف فهي دائما جزء من الكلمة، والاستثناء الوحيد هو كلمة (أفبالباطل) في القرآن الكريم.

٥- وإذا لم تنطبق أي من الحالات المذكورة أعلاه تصنف ال أداةَ تعريف.

تصنيف الحروف العربيّة

قُسّمت الحروف العربيّة في هذا المنهج إلى ثماني مجموعات: الأولى احتوت على جميع الحروف الأصلية التي من الممكن أن تكوّن الكلمة، والثانية والثالثة تحتوي الحروف التي لا يمكن أن تأتي سوابق، والرابعة والخامسة: الحروف التي لا يمكن أن تأتي بمنتصف الكلمة، والسادسة والسابعة: الحروف التي لا يمكن أن تأتي لواحق، والثامنة احتوت على مجموعة الحروف التي تأتي دائماً على شكل لواصق. واستُخدمت هذه التصنيفات لتطوير خوارزمية استخراج الجذور.

التقييم

استخدمت كلمات القرآن الكريم في التقييم، وتم إدخالها إلى نموذج المعالجة للقيام بما يأتي:

- حذف جميع الأرقام والرموز وعلامات الترقيم، وجميع العلامات، والكلمات المستبعدة، والحركات باستثناء علامة الشدة ( ).
- تقسيم النص إلى وحدات معجمية (**Tokens**).
- استثناء الكلمات المستبعدة (**Stop words**).
- حذف الكلمات المكررة.
- حفظ الكلمات المتبقية في ملف منفصل.
- أسفرت مخرجات الملف عن وجود ١٤٠٦٧ كلمة دون تكرار، وتراوح طول الكلمة بين حرفين إلى ثلاثة عشر حرفاً.

- وقُيّمت الخوارزمية على مرحلتين، هي:
- تقييم قواعد تمييز المقطع (ال).
- تقييم تصنيف الحروف العربيّة.

هناك ١٩٣٧ كلمة احتوت على (ال) من أصل ١٤٠٦٧ كلمة، إما على شكل (ال) التعريف مثل كلمة «والنبيين»، أو أصلية مثل كلمة «والد». وكانت الخوارزمية قادرة على تمييز (ال) بشكل صحيح ل ١٩٣٢ كلمة بنسبة (٩٩,٧٤٪)، وهناك فقط ٥ كلمات حُللت بشكل خاطئ.

اختبار تصنيف الحروف.

نجح النظام في تحليل ١٣٨٥٦ كلمة وفشل في ٢١١، فحروف الجذور المولّدة قورنت بالجذور المخزنة في قائمة الجذور. باستخدام هذه الطريقة، اعتبرت ١٣١٩٣ نتيجة صحيحة (٩٣،٧٩٪) و٦٦٣ نتيجة اعتبرت خاطئة (٤,٧١٪).

الخلاصة والتوصيات

للإجابة على سؤال الدراسة الذي ذكر في بدايتها، تم تحليل لغوي عميق لأوزان الكلمات ولواصقها في العربيّة. ونتيجة لذلك، بينت المرحلة الأولى من هذا المنهج الجديد أنه يمكننا إيجاد علاقات بين الحرف ومواضعها في الكلمة. وأن نتيجة التقييم تدل على ظهور خوارزمية واعدة لاستخراج الجذور. كما أن المنهج المفترض ما يزال حديثاً، وأظهر نتائج مميزة في حصر جذور الكلمات، ويمكن تحصيل نتائج أفضل إذا تم وضع المزيد من العلاقات تحت التجربة.

وخطتنا المستقبلية هي تطوير خوارزمية لتغطية العلاقات بين الحروف العربيّة لحصر الكلمات غير المنتظمة كالكلمات ثنائية الحروف واستخدام ذخيرة أكبر لفحص دقة النتائج.

Proceedings of the 16th conference on Computational linguistics - Volume 1 (COLING '96), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 89-94.

Year of publication: *1996*

Arabic Finite-State Morphological Analysis and Generation

Kenneth R. Beesley

التحليل والتوليد الصرفي العربي المحدود الحالة

ملخص

وصفت هذه الدراسة نظاماً ذا تغطية واسعة للقيام بتحليل صرفي (morphological analysis) وتوليد كلمات عربيّة، تكون مرسومة وفق قواعد الإملاء المعيارية (standard orthography)، سواء كانت مشكولة كلياً (fully voweled)، أو مشكولة جزئياً (partially voweled)، أو غير مشكولة (unvoweled). وتظهر التحليلات الجذر (root)، والوزن (pattern)، وجميع الزوائد (affixes)، ومع وسوم لخصائص تمثّل أقسام الكلام (part of speech)، والإسناد (Person)، والعدد (number)، والحالة الإعرابية للفعل (mood)، والمبني للمعلوم أو المجهول (voice)، وحالة الفعل (aspect)، الخ. بُني النظام باستخدام معجم وقواعد لغوية مستوحاة من نظام KIMMO للتحليل الصرفي، كما استُخدمت أدوات Xerox للتحليل الصرف ي للحالة المنتهية (finite state morphological analysis)، ونتج عن هذا محول عربي معجمي للحالات المنتهية (Finite-State Lexical Transducer) يعمل آنيا بنفس زمن التشغيل (runtime) المستخدم في الإنجليزية ولغات أخرى.

الأهداف

يجب على المحلل الصرفي العربي أن يتصف بما يلي:

- يجب أن يتعامل النظام مع نظام إملائي عربي حقيقي عبر الإنترنت وفق معايير عالمية، مثل (**ASMO449**)، و(**ISO8859-6**). كما أنه من الممكن ابتكار نظام دقيق لتحويل الحروف العربيّة إلى حروف لاتينية.
- يجب أن تحلل الكلمات العربيّة كما تظهر في النصوص الحقيقة. ذلك يعني أنه يمكن أن تكون الكلمة مشكولة كلياً، أو جزئياً أو غير مشكولة.
- يسهل البحث عن الكلمات في المعاجم المطبوعة والإلكترونية.
- يجب أن يكون النظام كبيراً وغير منتهٍ. فكل جذر يبرمج بتحديد الأوزان المناسبة له.
- يجب أن يكون النظام دقيقاً وفعّالاً، ويحلل بنجاح مئات أو آلاف الكلمات في الثانية الواحدة باستخدام أجهز الحاسوب الشائعة الاستخدام.
- يجب على النظام، بشكل فعّال ودقيق، أن يولد كلمات صحيحة إذا زُوِّدَ بجذر وأوسمة لخائص صرفية ذات صلة، ويجب أن يكون التحليل والتوليد الصرفي عمليتين عكسيتين ومباشرتين.

التطبيق

بداية يتم تحويل معاجم **ALPNET** إلى تنسيق **lexc**. وقد لوحظ أن تشابك الجذور والأوزان للغات السامية يمثّل عملية تقاطع، وهي عملية مدعومة في نظام (**Xerox**). وعليه إذا مُثّل أيّ حرف بعلامة السؤال ؟، والحروف الأصلية ب **C**، فإنه يكمن ترجمة الجذر (درس) الى (**?d?*r?*s***). وتقاطع الجذر مع الوزن (**CaCaCa**) ينتج الكلمة المجردة (الجذع) دَرَسَ.

وباستخدام العمليات النموذجية المتاحة خلال المترجم اللغوي وأدوات الحالة المنتهية (**finite state tool**) الأخرى، يمكن بناء التحليل طبقاً لاحتياجات اللغويين. ولأنّ السلسلة

العليا تسترجع نتيجةً للتحليل، فمن المهم تعريف السلسلة العليا كجذر يتبع بها مجموعة الوسوم الرمزية (tags) لتمثيل الخصائص الصرف نحوية (morphosyntactic features) البارزة في التحليل. فمثلاً، الكلمة «دَرَسَ» هي الصيغة الأولى للفعل التام المجرد المبني للمعلوم، وقد نتجت من الجذر (د ر س) والوزن (CVCVCV)، والتشكيل (aa) الذي يمثّل المبني للمعلوم.

بعد تشكيل المحولات المناسبة، تختفي المستويات المتوسطة في التحويل منتجةً ربطاً مباشراً بين المستويات العليا والمستويات الدنيا. وتسمى عملية إنتاج المحول الفردي «المحول المعجمي» (lexicon transducer). جميع الجذوع (الكلمات المجردة) وعددها ٨٥٠٠٠ أضيفت آليا في عملية التحليل، السوابق واللواحق المناسبة وجدت في المحولات المعجمية، وأضيفت بالطريقة الإلصاقية العادية.

وأثناء التشغيل تم توصيل جميع السلاسل المحللة بمسارات داخلية خلال المولد اللغوي، في حين تكون سلاسل النتائج في القمة، ومثل جميعها محولات الحالة المنتهية، تتولد كلمات محللة دقيقة لغوياً. الكلمة العربيّة المفردة يمكن أن تهجى بأوجه عديدة وتعتمد على كيفية القراءة، لذلك النظام يبين جميع الأشكال الممكنة للكلمة مع التشكيل الممكن. ويمكن للتوليد في النظام أن يكون كامل التشكيل أو النموذج المشكول جزئياً والنموذج غير المشكول.

Proceedings of the 2015 International Conference on Soft Computing and Software Engineering (SCSE'15), Procedia Computer Science, Volume 62, 2015, Pages 521-528

Year of publication: *2015*

Morphological Analysis in the Environment "TELA"

Khaireddine Bachaa, Mounir Zriguib

التحليل الصرفي في بيئة تعليمية للغة العربيّة بمساعدة الحاسوب المسماة "تيلا"

لا تزال الحاجة إلى معالجة اللغات الطبيعية التلقائية في الموارد المعجمية الكبيرة في نمو مستمر، ويجب أن تُعد إدارة هذه المعرفة أولوية لأنها عنصر أساسي في كفاءة تطبيقات معالجة اللغات الطبيعية التي تستخدم تلك الموارد، وبالتالي تزيد من الاهتمام بتطور قواعد البيانات المعجمية القابلة لإعادة الاستخدام والمستقلة عن تطبيق لغة معينة.

لقد ركز هذا البحث على دراسة اللغة العربيّة، فعلى الرغم من موقعها لغةً خامسةً في العالم، وأن هناك أكثر من(٥٠٠٠٠) موقع عربي على شبكة الإنترنت وأكثر من ٣٢٠ مليون شخص يتكلمها، على الرغم من كل ذلك، لا يوجد محلل قادر على التعامل بقوة مع جميع الظواهر الصرف-نحوية. ومع ذلك، يلاحظ أن هناك زيادة في المحتوى النصي باللغة العربيّة، وخاصة على شبكة الإنترنت. إلى يومنا هذا، لا تزال معالجة موارد المعلومات وتشغيلها تشكل تحدياً للباحثين في مجال المعالجة الآلية للغات الطبيعية. وفي هذا السياق، نوقشت مشكلة التحليل الصرف-نحوي التلقائي الدقيق للغة العربيّة، كما بدأ الباحثان في بناء مجموعة من الموارد اللغوية للغة العربيّة من أجل بناء نموذج آلي.

في هذا البحث، اقترح الباحثان نهجاً إحصائياً لبناء محلل صرفي سمي (**TELA-MA**) ويقصد به « التوجه نحو بيئة تعليمية للتحليل الصرفي للغة العربيّة»، حيث وُصفت التقنية العامة والموارد المنفذة لإنجاز مهمة بناء المحلل. هذا العمل يشرح بالتفصيل التجارب المختلفة

وهي: تحليل ذخيرة التدريب (Training Corpus) واختيار مجموعة الوسوم. وأخيراً، تقييم نظام التجزئة (Segmentation System) الذي ينهي هذه المرحلة من العمل على التحليل الصرفي للكلمات العربيّة، وقد كانت النتائج مشجعة.

يتكوّن محلل (TELA-MA) من جزأين رئيسيين: الجزء الأول يتعلق بالتجزئة المعجمية للنص العربي؛ ويتم ذلك عن طريق التجزئة إلى فقرات أو جمل أو كلمات. أما الجزء الثاني فينقسم إلى مرحلتين هما مرحلة التحليل الصرفي المعتمدة على آلات الحالة المنتهية (Finite state automata)، ومرحلة التوسيم (Labeling Phase). ومن أجل إنجاز محلل صرفي، فقد أشار الباحثان إلى بعض أنظمة التحليل الصرفي للغة العربيّة، وأوضحا مساوئ كل واحد منها ومبادئ التشغيل المستخدمة في محاولة للتغلب على تلك المساوئ والحصول على تطبيق يسعى إلى تحسين المعالجة باستخدام قواعد صرفية واستجابات المتعلمين في بيئة التعلّم للغة العربيّة بمساعدة الحاسوب"TELA" ، وكان من بين مميزات المحلل المقترح التجزئة المعجمية للنص العربي، حيث يستند نظام TELA-MA على مجموعة من القواعد الصرفية والمعجمية التي حددت من خلال عينة المعلمين، الذين يعدون متخصصين وخبراء في هذه العملية، وسُمّيت التجزئة الموجهة لإزالة الغموض الذي يؤثر على اللغة العربيّة، وقد عُرضت في ذخيرة التدريب. ومن ناحية أخرى هناك عينة الطلبة الذين شاركوا في تنظيم التطبيق، وحقق النموذج أداءً جيدا بنسبة ٨٩٪..

CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning, pages 135–142, Manchester, August 2008

Year of publication: 2008

Context-based Arabic Morphological Analysis for Machine Translation

ThuyLinh Nguyen, Stephan Vogel

التحليل الصرفي العربي المعتمد على السياق لغرض الترجمة الآلية

في هذه الورقة نقدّم تقنية جديدة للتحليل الصرفي معالجةً أوليةً لغرض الترجمة الآلية من اللغة العربيّة إلى اللغة الإنجليزية. تعتمد الترجمة الآلية الإحصائية (**SMT statistical machine translation**) بشكل كبير على نموذج محاذاة الكلمات (**word alignment model**) بين اللغة الأولى (**source langaue**) واللغة المستهدفة (**target language**). لكن هناك عدم تطابق بين لغة ذات بنية صرفية غنية مثل اللغة العربيّة، ولغة ذات بنية صرفية فقيرة مثل اللغة الإنجليزية. إنّ كلمة واحدة في اللغة العربيّة عادةً ما تقابل عديدة كلمات في اللغة الإنجليزية.

الطريقة

نقوم أولاً بالمعالجة المسبقة لذخيرة التدريب (**training corpus**) وذلك بتقطيع الكلمات فيها الى سلاسل من وحدات صرفية (**morphemes**) على الشكل سوابق وجذع ولواحق، مع إسقاط بعض اللواحق كالعلامة الإعرابية (**case marker**) لعدم وجود ما يقابلها في اللغة الإنجليزية. والتنيجة من هذه العملية هي ذخيرة محللة صرفياً بشكل كامل. وبعد حذف العلامات الإعرابية بلغ عدد نماذج العينات الصرفية المحللة باللغة العربيّة (١,٧) ضعفاً بالمقارنة مع عدد الكلمات باللغة الإنجليزية. ولوحظ أن الترجمة من اللغة العربيّة ينتج عنها الكثير من الكلمات غير المعروفة (بدون ترجمة) (**unknown word**) في عينة الفحص (**test set**) مما يؤدي إلى تقابل عدة كلمات إنجليزية مع كلمة عربيّة واحدة. أما على مستوى

الوحدة الصرفية (morpheme)، تحُاذي الكلمة الإنجليزية وحدة صرفية واحدة في الذخيرة المحللة صرفياً، ولكن بعض السوابق واللواحق في الذخيرة المحللة صرفياً قد لا تحاذي أية كلمة إنجليزية مطلقاً. إنّ استخدام الذخيرة المحللة صرفياً للترجمة قد يؤدي إلى تقديم وحدات صرفية زائدة في جانب اللغة الأولى. والهدف من استخدام الذخيرة المحللة صرفياً للترجمة الآلية هو إزالة السوابق واللواحق غير المتحاذية (nonaligned) باستخدام منهجية مشتقة من البيانات (data-driven approach).

طُبّقت في دراسة سابقة نموذج القناة المصدرية (source-channel model) لحل مشكلة الغموض الصرفي، بحيث يكون النموذج المصدري نموذجاً موحداً يعرّف مجموعة التحليل. واستخدمت هذه الدراسة أداة (BAMA) المحلل الصرفي، التي وفّرت مجموعة التحليل للبدائل الصرفية.

لغرض استخراج الوحدات الصرفية العربيّة التي تحاذي النص الإنجليزي، استُخدم أداة (GIZA++). واختير النموذجان (IBM٣ و IBM٤)، وعند اعتماد الإنجليزية بنيةً مصدرية (لغة أولى)، أجبرت النماذج الموافقة الكلمات الإنجليزية لتوليد موافقاتها في الوحدة الصرفية العربيّة.

تقليص (reducing) نموذج النظام الصرفي

تعدّ الذخيرة الصرفية المقلصة الخيار الأفضل للتحليل الصرفي للترجمة الآلية، لأنه من غير الممكن وسم الوحدات الصرفية لجملة الفحص (test sentence) دون مرجعية باللغة الإنجليزية، ونحتاج لتعلم نموذج لتوسيم الوحدة الصرفية(morpheme tagging model) الذي يقوم بتوقع توزيع السلاسل الموسومة التي تعطي جملة محللة صرفياً باستخدام الخطوات المذكورة أعلاه.

نتائج التجربة

طُبّقت تقنية التحليل الصرفي على ذخيرة لغوية احتوت على ٥,٢ ملايين كلمة عربيّة

(١٧٧ ألف جملة) في مجال الأخبار و ١٥٩ ألف كلمة عربيّة (٢٠ ألف جملة) في محاورات السفر وقد نتجت تحسينات في الترجمة في كلا التجربتين.

إحدى المزايا الواضحة في استخدام الترجمة المعتمدة على التحليل الصرفي خلال ترجمة الكلمة الأصلية هو تقليل عدد الكلمات غير المترجمة.

الخلاصة والتوصيات

تم تنفيذ تقنية معالجة صرفية مسبقة للترجمة من العربيّة إلى الإنجليزية، وتفوّقت هذه التقنية على الأنظمة العربيّة السابقة عند تطبيقه على بيانات ذات حجم صغير ومتوسط. في البحوث المستقبليّة نخطط لتطوير النموذج ليدمج الوحدات الصرفية ونخطط أيضاً لدراسة تأثير أطوال العبارات ودلالتها.

٣-٢-٢ النحو

أبحاث النحو من أهم الأبحاث في معالجة اللغة العربية، وتحتوي على تسعة أبحاث: ثلاثة منها من النوع أ، تضمنت: دراسة مسحية للنحو المعجمي الوظيفي في السياق العربي، والوصف الصوري للهيكل النحوي العربي في إطار نظرية الترابط والحوكمة، والمحلل النحوي العربي المسمى "بالعربي".

أما الأبحاث من نوع ب فهي: اللُغة العَربيّة بين الصياغة المنطقية الرياضية والحَوسَبة ، ومُحلّل نحوي قائم على محلل شبكة التحوّلات المتكررة للُغة العربيّة، وإزالة الغموض من النص العربي باستعمال نحو الاعتماد، وتحليل مقارن لأنظمة الفعل العربية والإنجليزية باستعمال الذخيرة اللغوية للقرآن الكريم ، وأشجار التحليل للجمل العربية باستخدام حزمة برمجيات اللغة الطبيعية، والنهج القائم على القواعد في معالجة اللغة العربية الطبيعية.

International Journal of Computing and Network Technology (nt. J. Com. Net. Tech. 4, No. 3141-146 , (Sep-2016)

Year of publication: 2016

A Survey of Lexical Functional Grammar in the Arabic Context

Said A. Salloum, Mostafa Al-Emran, Khaled Shaalan

دراسة مسحية للنحو المعجمي الوظيفي في السياق العربي

النحو المعجمي الوظيفي (**Lexical Functional Grammar**) هو فرضية لغوية ابتكرتها جوان بريسنان [**Joan Bresnan**] في عام ١٩٧٠. ويلعب النحو المعجمي الوظيفي دورا مهما في المعالجة االحاسوبية للغات الطبيعية لأنه يوفر بنية لوصف المعلومات النحوية لأي لغة طبيعية. ويميز النحو المعجمي الوظيفي بين مستويين من الوصف لكل جملة في اللغة، وهما: صيغة الشجرة (البنية التأسيسية) (**c-structure**)، والبنية الوظيفية (**f-structure**) التي تمثل الوظائف النحوية مثل الفاعل والمفعول به والعلاقة بينهما على صورة مصفوفات تعرض الكلمة والوظيفة.

فك اللبس في الجملة العربية:

أشارت الدراسات إلى أن عدم استخدام التشكيل عند كتابة الجمل العربية في الغالب يزيد من غموض اللغة، ويبطئ من تطور معالجتها. وفي حال فك ذلك الغموض فإن نطاق التفسيرات الممكنة يصبح أقل وتصبح اللغة أكثر وضوحاً. إن نظام التحليل النحوي المطوّر للغة العربية يتضمن ثلاثة عناصر: المعجم، والمحلل الصرفي، والمحلل النحوي. ونتيجة لتطبيق نهج فك اللبس الذي يعتمد على التحليل الصرفي والنحوي، فإن المحلل الصرفي يعطي كل القراءات الممكنة لكلمة عربية معينة، وبعدها يأتي دور المحلل النحوي الذي يقوم بالتحليل بتطبيق القواعد النحوية للغة، مما يقلل عدد الاحتمالات الممكنة ويزيد من وضوح الجملة وفك اللبس فيها.

قواعد الجمل المعرفة

أوضحت بعض الدراسات أن اللغة العربية تختلف عن اللغات الأخرى طريقتها الوصفية في التعبير عن قواعد النحو في أنها يُعبّر عن قواعد النحو فيها بطريقة وصفية. وقد كانت هناك بعض الجهود التي حاولت صياغة شكل بنيوي للجمل العربية، مثل نموذج النحو المعجمي الوظيفي، وقواعد التبعية، والقواعد الوظيفية، لكن تلك المسألة ما زالت قيد البحث والنقاش. وقد طُوّر وصف بنيوي للنحو العربي من بعض الباحثين يسمى نحو الجمل المقيدة (Definite Clause Grammar DCG) باستخدام لغة (prolog)، واستخدم في هذا الوصف في المحلل الصرفي.

ركزت هذه الدراسة على تصميم محلل نحوي للغة العربية وتطبيقه عليها، ليكون جزءاً من نظام ترجمة آلية. واستخدم فيه (نحو الجمل المقيدة). وقد طُوِّر هذا النظام على مرحلتين: المرحلة الأولى تتمثل بالحصول على القواعد التي تشكل نحو اللغة العربية، وتعطي وصفاً دقيقاً لماهية الجمل الصحيحة نحوياً. وكانت الجمل مستقاة من وثائق موضوعها الزراعة. أما المرحلة الثانية فهي التطبيق الفعلي للمحلل النحوي، مع تعيين الهيكل أو البنية النحوية للجمل المدخلة، حيث يقوم المحلل بترميز القواعد النحوية للغة العربية، والتأثير الحاصل على مكونات الجملة عند تطبيق تلك القواعد. وقد بني هذا المحلل باعتباره نموذجا يمكن تطبيقه مع أي نظام آلي يتعلق بالتحليل النحوي.

وعند تصميم هذا المحلل تجنب الباحثون مشاكل الغموض واللبس قدر المستطاع، فعملية فك اللبس الصرفي لا يمكن أن تكون شاملة، وهذا يشكل عائقاً بحد ذاته. وبإجراء التجارب على نصوص ووثائق حقيقية كانت النتائج مرضية. وقدمت إحدى الدراسات آلية جديدة تستخدم بشكل أساسي ثلاثة محللات، بناء على طريقتين رئيسيتين: تبديل المحلل وتهجين المحلل، وذلك لتحقيق نتائج جديدة بدقة تحليلية عالية لتقليل أخطاء التحليل مقارنة بالتقنيات الأخرى. وقد جُرّبت تلك المحللات الثلاث المركبة بتدريبها على أشجار التحليل. وألقت دراسة أخرى الضوء على تطوير محلل يستخدم جملا عربية فصحى. كما أن هذا المحلل المطور يستخدم النحو

المقيد لتقليل اللبس عن طريق توظيف بعض سمات الدلالات المعجمية التي تستخدم بشكل أساسي لفك بنية الجملة الغامضة، واستخدمت لغة برولوغ لتطبيق هذا المحلل.

واستخدم باحثون آخرون طريقتين للتحليل النحوي العربي عن طريق المحللات المعتمدة على أشجار التحليل والنحو المعجمي الوظيفي الأوتوماتيكي الذي يطبق مقياس دقة الاختبار (ف) [F-score]. وقد استخدمت مدونة بن العربية المشجرة (PATB) ميزة لوغاريثمات وسم اللغة العربية (A3) التي تستفيد من الوسومات الوظيفية لتضمين مقياس دقة الاختبار (ف) مع شجرة التحليل.

وقد أجري تغيير فعال على محلل بيكل (Dan Bikel's analyzer) من أجل دمج مجموعة العبارات ومعرفة السمات الوظيفية في مدونة بن العربية المشجرة (PATB) عن طريق اختيار البيانات التدريبية التي تغطي السمات الوظيفية. وبعد تطبيق هذه التقنية، أظهرت النتيجة درجة من الدقة وصلت إلى نسبة ٧٧٪.

وقد وضعت دراسة أخرى منهجية لإنتاج أشجار تحليل للجمل العربية وتحديداً الجمل القرآنية المقترحة من خلال استخدام أداة معالجة اللغة الطبيعية NLTK [Natural Language Toolkit] بحيث يتم خلال تلك العملية بناء قاموس و قواعد نحوية بدون سياق والاستخدام المتكرر لمحلل NLTK . وتعد الأشجار النحوية الناتجة مكونات لأشجار التحليل. ويمكن استخدام هذا النهج أيضا في العديد من التطبيقات، لأن دمج محلل صرفي ومحلل نحوي يؤدي ببساطة إلى أتمتة العملية. وقد تم تطبيق خوارزمية تحليل من أعلى إلى أسفل لتحليل الجمل العربية من خلال استخدام NLTK.

ووضعت دراسة حديثة أخرى نظاما يحلل الجمل العربية الفصحى بواسطة استخدام مصادر أشجار التحليل. يتّخذ هذا النظام المطوّر الجملة العربية مدخلا، وينتج شجرة التحليل لتلك الجملة بناء على النموذج المبني الذي أنشئ في مرحلة تدريب البيانات، وأشارت نتائج التجارب إلى أن النظام حقق درجة من الصحة (accuracy) وصلت إلى ٨٢,٤٪ ونسبة ٨٦,٦٪ للاسترجاع (recall) و٨٤,٤٪ لمقياس الدقة (precision).

هياكل وبنى التبعية للنحو المعجمي الوظيفي العربي

أظهرت إحدى الدراسات أن المحللات الإحصائية في الأبحاث المتعلقة بأشجار التحليل النحوي تحتاج مزيدا من الدعم أكثر من تلك التي تستخدم قواعد اللغة يدويا. ولكن لوحظ أن نقاط الضعف لتلك المحللات تكمن في عدم قدرتها على تمييز الخصائص النحوية والعملية التي تتطلبها بعض التطبيقات الخاصة بالدلالة. كما لوحظ أن أهمية تزويد المحللات الضعيفة ببيانات عميقة مساندة تشكل نقاط ضعف لتلك المحللات. ويمكن استخدام المعلومات التي تم ترميزها في أشجار التحليل لمدونة بن العربية (Penn-II Treebank PTB) التي أزيلت منها الموارد النحوية واستخدمت في كل من التوليد والتحليل أوتوماتيكيا لوسم كل عقدة في الشجرة جنبا إلى جنب مع مقياس دقة الاختبار (ف). وقد طُبّق هذا النهج على لغات مختلفة مثل: الألمانية والصينية والإسبانية والفرنسية. علاوة على ذلك، تم تطبيق اكتساب النحو المعجمي الوظيفي على اللغة العربية وعلى أشجار التحليل العربية (Penn Arabic Treebank). وكشفت النتائج أن معظم أشجار (ATB) حققت مقياس دقة الاختبار (ف) بشكل كامل. وقد أجري التقييم النوعي باستخدام مجموعة المعايير الذهبية وحققت درجة دقة بنسبة ٩٥٪.

مدونة أشجار بن العربية ((Penn Arabic Treebank (ATB)

بدأت دراسة أشجار التحليل العربية عام ٢٠٠١ بهدف شرح الأخبار العربية، واحتوت تلك الأشجار على ٢٣٦١١ جملة. يستخدم وسم أشجار بن العربية (ATB) مفهوم العناقيد الفارغة لتمييز التوابع الطويلة مثل جمل الوصل والأسئلة، حيث وسمت العناقيد الفارغة بعد الفعل بوسم وظيفي (SBJ) ، فاللغة العربية من اللغات التي تسمح بأن يكون موقع الفاعل فارغاً.

والخلاصة أن للنحو المعجمي الوظيفي (LFG) تأثيرا كبيرا على حل مشكلات معالجة اللغة الطبيعية (NLP). وعرضت الدراسة تحليل دراسات بحثية ركزت على استخدامه لحل الغموض في اللغة العربية وفي تحليل الجمل العربية.

من المتوقع أن يحقق نظام المحلل العربي لتحليل النصوص العلمية الحديثة باستخدام النحو المعجمي الوظيفي نتائج مرضية. لكن ذلك يتطلب إنشاء كمية هائلة من أشجار التحليل بسبب مشكلات الغموض. أما بالنسبة للمستقبل، فالباحثون يعملون على حل مشكلات الغموض والحصول على شجرة تحليل محددة بدلاً من عدة قراءات محتملة. وقد يركز البحث المستقبلي على تطوير محلل عربي يأخذ جملة عربية باعتبارها مدخلا، وينتج الشجرة المقابلة في تنسيق واجهة المستخدم الرسومية.

Computacion y Sistemas Vol. 18, No. 32014 , pp. 611-625

Year of publication: 2014

Formal Description of Arabic Syntactic Structure in the Framework of the Government and Binding Theory

Bassam Hammo, Asma Moubaiddin, Nadim Obeid, Abeer Tuffaha

الوصف الصوري للهيكل النحوي العربي في إطار نظرية الترابط والحوكمة

المقدمة

كثير من الكلمات تنقل المعنى، ولكن عندما تجمع معا على أساس البنية النحوية فإنها تنقل معاني أكبر؛ وبالتالي فإن تحديد البنية (بناء الجملة) هو الخطوة الأولى نحو فهم معنى الجملة. والتحليل النحوي (التحليل) هو إجراء يتعرف على الجملة ويكتشف كيف بنيت (أي يعطي هيكلها النحوي). أما الإعراب فيتضمن معرفة ما إذا كانت الجملة تنتمي إلى لغة معينة، أي فيما إذا كانت تتبع جميع القواعد التي تنص عليها هذه اللغة. واكتشاف البنية (التحليل) يتضمن تحديد علامات ووضعها على مكونات مختلفة من الجملة (أي العبارات والأجزاء الفردية من الكلام مثل الاسم، الفعل، حرف الجر، وما إلى ذلك) .

نظرية الترابط والحوكمة (Government- Binding Theory)

نظرية الترابط والحوكمة (Government- Binding Theory) (غب) (GB) هي طريقة عالمية للقواعد، تحتوي على قواعد ومبادئ يمكن أن تطبق في كل اللغات، فهي تحتوي على قواعد عالمية، ومع ذلك، هناك الكثير من الاختلافات بين اللغات مثل اختلاف تسلسل الفاعل والفعل والمفعول به. ومن المتفق عليه أن كل لغة لها ترتيب معين للكلمات الأساسية، وأن كل تسلسل آخر ينتج عن تحريك مكونات الجملة وفق قواعد معينة. تُنظّم الكلمات بشكل هرمي

في وحدات أكبر تسمى العبارات. وتشمل مكونات العبارة ما يلي:

١- العبارة الانعكاسية (Inflectional Phrase) : عبارة العطف، وتتكون من عناصر تتضمن العدد، ونوع الجنس .

٢- العبارة المكملة (Complementizer Phrase) : عبارة تبتدئ بعبارة انعكاسية.

٣- الجملة الاسمية (Noun Phrase) : جملة تبدأ باسم.

٤- الجملة الفعلية (Verb Phrase): جملة تبدأ بفعل.

٥- عبارة الصفة: (Adjectival Phrase) وتبتدئ بصفة.

٦- عبارة حرف الجار والمجرور (Prepositional Phrase) وتبتدئ بحرف الجر.

أجريت محاولة لتطبيق نظرية الترابط والحوكمة على بعض الجمل البسيطة في اللغة العربية وخاصة:

١- الجمل التي تليها أشباه جمل مثل الجار والمجرور (Prepositional Phrase) .

٢- أشباه الجملة المصدرية (Complementizer Phrase) .

المحلل النحوي العربي (The Arabic parser)

يأخذ المحلل النحوي الجملة باعتبارها مدخلات ويحللها ليحدد إذا كانت الجملة صحيحة من الناحية التركيبية، ويقوم كذلك بتمثيل تركيبها باستخدام الأقواس. ومن أجل إضافة أقسام الكلام إلى الجملة، استخدمنا برنامج أرامورف، وهو محلل باك والتر الصرفي للغة العربية والمطوّر بلغة البرمجة جافا.

القواعد النحوية (Grammar Rules)

تتضمن القواعد النحوية مجموعة من قواعد اللغة العربية مستنبطة من تحليل الجمل العربية باستخدام نظرية غب. وهي مقسمة إلى قسمين: يتضمن القسم الأول القواعد النحوية التي لا

تهتم بعلامات التشكيل، أما القسم الثاني فيتضمن قواعد الجمل مع مراعاة الحالة. وفي كلتا الحالتين، يتم سرد الجملة بمكوناتها جملة جملة، تليها قواعد بناء جملة من الجمل التي تم تحليلها.

تنفيذ النظام والنتائج

يتم تمثيل جملة الإدخال إلى المحلل النحوي باعتبارها مجموعة متسلسلة من الوسوم المستخدمة لوسم أقسام الكلام في الجملة. ويأخذ المحلل النحوي تسلسل علامات الجملة الموسومة ويعطي بناء (نواتج) للتحقق من سلامة تركيب الجملة. وقد اعتمدنا نهجاً متكرراً لمسح تركيب الجملة من الأعلى إلى ألاسفل؛ وتراكيب قواعدية تبدأ من مستوى الجملة (S)، والاستمرار في إيجاد قواعد للعبارات (المستوى المتوسط) وأخيرا الانتهاء بتمثيل أقسام الكلام (أدنى مستوى).

لقد اخترنا استخدام لغة البرمجة « برولوغ» لتنفيذ المحلل النحوي، ويمكن استخدام هذه اللغة بشكل فعال في تحليل اللغة الطبيعية للأسباب التالية: (١) برولوغ هي لغة البرمجة المنطقية، وتبدو مناسبة للتعبير عن قواعد النحو و (٢) لم نكن نهدف إلى اختبار فعالية قواعد اللغة العربية، لقد استخدمنا ملفين: يتم استخدام الأول من المستخدم عندما لا يتطلب تحليل الجملة أخذ الحركات في عين الاعتبار، أما الملف الثاني فيستخدم عندما ينظر المستخدم الى الحركات.

مقارنة مع الأعمال السابقة

بقدر ما نعلم فإن هناك محاولات قليلة جدا لتطوير محلل «غب» للغة العربية. وقد أجريت بعض محاولات التحليل النحوي العربي استنادا إلى معجم من الكلمات متضمن على خصائص الكلمات المعجمية والنحوية التي تساعد في إزالة الغموض في بنية الجملة. وتستخدم هذه الطرق محللاً صرفياً، وتقوم على تحديد عدد من القواعد على المستوى التصنيفي للكلمات مثل: الفاعل، المفعول به، الخ. ويقسم المعجم إلى ثلاث فئات: الأسماء والأفعال والحروف. وترتبط مدخلات المعاجم بنوعين من الخصائص: التركيب النحوي والدلالة اللفظية. وتستخدم الخصائص النحويّة لحل الغموض النحوي مثل زمن الفعل، الفاعل والمفعول به والجنس والعدد. وتستخدم الخصائص الدلالية لحل الغموض المعجمي.

الخلاصة والأعمال المستقبلية

إن التحليل العربي النحوي القائم على الأتمته هو مجال بحثي واسع النطاق نظرا لثراء اللغة العربية. في هذه الورقة، قمنا بتحليل البنية النحوية لبعض الجمل العربية البسيطة على أساس نظرية «غب». وقمنا بدراسة الترتيب لكلمات مختلفة باللغة العربية ودرسنا اشتقاقها. وقمنا أيضا بتحليل عدة تراكيب تشمل: VSO,VOS,SVO ، الجمل الاسمية، الجمل الاسمية مع إن وأخواتها وجمل السؤال. واستخدمنا التحليل لتطوير قواعد نحوية لجزء من قواعد اللغة العربية، ووضعنا مجموعتين من القواعد: (١) قواعد تراكيب الجمل التي لا تأخذ الحركات في الاعتبار و (٢) قواعد تراكيب الجمل مع اعتبار الحركات.

أظهرت النتائج دقة عالية عندما تم وسم أقسام الكلام في الجمل التي تأخذ الحركات في عين الاعتبار. ومن المهم أن نلاحظ أن النظام لم يكتمل بعد، ونحن نعمل على اختباره باستخدام ذخيرة لغوية قياسية، ومن ثم مقارنة النتائج مع أنظمة مماثلة. ويمتاز النظام المقترح بالمرونة ويمكن تطويره بحيث يمكن إضافة مزيد من التعديلات عليه. ويحدونا الأمل في تعزيز النظام من خلال (١) استخدام محلل صرفي من شأنه أن يوفر معلومات هامة حول الكلمات مثل تحديد اللواصق، والعدد والجنس و (٢) من خلال إضافة المزيد من القواعد للتعامل مع المزيد من تراكيب الجمل.

International Journal of Social Science and Humanity vol. 6, no. 5, pp. 341-3465 , May 2016

Year of publication: 2016

Bel-Arabi: Advanced Arabic Grammar Analyzer

Michael Nawar Ibrahim, Mahmoud N. Mahmoud, Dina A. El-Reedy

"بالعربي" – محلل النحو العربي المتقدم

يهدف هذا البحث إلى بناء محلل نحوي آلي (معرب) لجمل اللغة العربية، وهذا يعد من أكثر العمليات تعقيداً في حقل معالجة اللغات الطبيعية، فبناء محلل قواعد عالي الدقة مستند على القواعد يعد أمراً معقداً، ولذا يقترح البحث نظامًا هجينًا بين الطرق المعتمدة على التعلّم، والطرق المعتمدة على القواعد، ما يعد بدقة مقبولة وسهولة في التطبيق.

الإعراب يعني معرفة الوصف النحوي لكل كلمة في الجملة، والنظام المقترح هنا يغطي القواعد الأساسية للجمل الاسمية والفعلية ضمن المحددات الآتية: كتابة الجملة بشكل صحيح صرفيا ونحويا، وأن تكون الأفعال الواردة في الجملة مبنية للمعلوم، وألا يكون للنظام علاقة بالمعنى أو بالاستخدام الخاطئ لمعاني الألفاظ الواردة في الجمل. وليس من السهل تقييم نظام (بالعربي) بسبب عدم وجود بيانات معيارية لتحليل القواعد العربية، لذا استخدمنا ٦٠٠ جملة لتقييم النظام.

عند النظر إلى أنظمة معالجة اللغة العربية الأخرى، نجد أن التركيز كان على التحليل الصرفي، وقد تحقق نجاح بعض الأنظمة، ومنها MADA و TOKAN وهي عبارة عن مجموعة من الأدوات من أجل فك اللبس الصرفي، ومعرفة أقسام الكلام (POS tagging)، والتشكيل، ومعرفة معاني الكلمات، وفروعها، وقد عمل النظامان على مجموعة من عمليات معالجة اللغة العربية، فكان MADA نظامًا للتحليل الصرفي وفك اللبس، أما TOKAN فهو (tokenizer) عام أي محلل معجمي عام للنص الموضح بنظام MADA، وعلى هذا فإن النظامين يقدمان معًا حلًا لعدد من مشكلات اللغة العربية.

وهناك نظام آخر لمعالجة مشكلات أخرى هو **AMIRA** وهو مجموعة أدوات للتحليل المعجمي (**tokenization**) للعربية، ومعرفة أقسام الكلام، وأنواع الكلمات في الجمل (**Base Phrase Chunking**)، والتعرف على أسماء الأشياء (**Named Entities Recognition**)، ويعتمد نظام **AMIRA** على التعلم المراقَب دون اعتماد صريح على معرفة عميقة بالصرف، وبالتالي يختلف عن أنظمة مثل **MADA** فهو يعتمد على المعلومات السطحية لتعلم التعميم، وفي العموم تعتمد أدواته على استخدام أنماط موحدة للتعامل مع كل مشكلة من مشاكل المحتوى كمشكلة تصنيف.

مراجع سابقة

اقترحت إحدى الدراسات نظاماً لأتمتة الإعراب للجمل العربية بشكل عام، ويفترض هذا النظام أن تكون الجمل المدخلة صحيحة صرفيا ونحويا وأنها مبنية للمعلوم. وقام باحث آخر بتجربة عدة منهجيات اللبس الصرفي والنحوي للعربية، فبنى محللاً عربياً يسمح بكتابة قواعد اللغة، واختبر نظامه على جمل قصيرة، وقال إن نسبة الدقة وصلت ٩٢٪. . وبنى باحث ثالث ما سمي ببنك الأشجار النحوية العربي لجامعة كولومبيا (**Columbia Arabic Treebank**) وهي قاعدة بيانات للتحليل النحوي للجمل العربية. وهي تمتاز عن قواعد البيانات المشجرة السابقة بتركيزها على السرعة مع وجود قيود على ثرائها اللغوي. وفيها فكرتان أساسيتان، وهما عدم التركيز على تكرار المعلومات، واستخدام مصطلحات مستلهمة من تركيب الجملة العربية التقليدي؛ لذا يتم التحليل بتطبيق طريقة إعراب بسيطة.

وفي بحث رابع بنيت قاعدة بيانات للقرآن الكريم سميت بنك الأشجار النحوية المعتمدة للقرآن الكريم (**Quranic Arabic Dependency Treebank**)، وتتكون من ٤٣٠ , ٧٧ كلمة من القرآن. ويختلف هذا المشروع عن غيره بتقديمه نموذجاً لغوياً حاسوبياً يستند على القواعد العربية التقليدية التاريخية.

معظم الأعمال التي استعرضناها في هذا البحث تركز على الجمل القصيرة، وتستخدم قواعد مبنية بشكل يدوي تتطلب وقتاً طويلاً لإنتاجها، ويصعب تطبيقها على بيانات غير مقيدة، كما استخدمت تقنيات إعراب تقليدية لجمل فعلية بسيطة أو جمل اسمية قصيرة.

منهجية البحث

يأخذ النظام المقترح جملة ويسم كل مقطع منها بعلامة تدل عليه كما يأتي:

العلامات العربية: فعل مضارع، فعل أمر، فعل ماض، فاعل، مفعول به، مفعول مطلق، نائب للمفعول المطلق، مبتدأ، خبر، مبتدأ مؤخر، اسم إن، خبر إن، اسم كان، خبر كان، اسم كاد، خبر كاد، بدل، نعت، توكيد معنوي، توكيد لفظي، معطوف، مضاف إليه، اسم مجرور، تمييز، مستثنى، منادى، ظرف، ضمير، حرف نفي، حرف نصب، حرف جزم، حرف جر، حرف استثناء، حرف عطف، أداة نداء، حرف تحقيق، حرف تقليل، علامة ترميز، حرف.

الحالات الإعرابية: مرفوع، منصوب، مجرور، مجزوم، مبني، حذف النون، حذف حرف العلة، الكسرة، الضمة، الفتحة، السكون، الواو والنون، الياء والنون، الألف والنون. ولمعرفة قسم الكلام (POS tag) ونوعه (BP chunk) لكل مقطع من الجملة وبنائها الشكلي، نستخدم النظام لنحدد العلامة والحالة وإشارة كل كلمة في الجملة.

ويمكن وصف مدخلات وعمل محلل القواعد كما يلي:

مدخلات: جملة عربية كاملة.

سياق: الجملة كاملة.

خصائص: استخلاص نحو كلمات الجملة، باستخدام مجذع الكلمات (Stemmer) ومحدد نوع الكلام (POS tagger) ومكدس العبارات (BP chunker) والمحلل صرفي (morphological analyzer) لاستخلاص خصائص صرفية إضافية للكلمات في الجملة.

مكونات النظام

* المحلل الصرفي (morphological analyzer): يحتوي على خصائص إضافية، مثل استخلاص أوزان الكلمات مثل: كاتب على وزن فاعل، كما قد تستخدم لاستخلاص الجذور ومعرفة المؤنث والمذكر والمفرد والجمع.

* مجذّع (Stemmer) لا بد من تقسيم سيل الحروف في نصوص اللغات الطبيعية إلى وحدات ذات معنى قبل معالجتها، والمجذع مسؤول عن تحديد حدود الكلمات، والسوابق واللواحق فيها.

* محدد أقسام الكلام (Part of Speech tagger) تمثل هذه العملية تحديد القسم من الكلام الذي تنتمي له كل كلمة بناء على معناها والسياق، وهناك صعوبتان في هذه المرحلة: الأولى هي عدم وضوح معاني الكلمات؛ فمعظم الكلمات في اللغة تنتمي لأكثر من قسم من أقسام الكلام، والثانية بسبب الكلمات غير المعروفة بالنسبة للنظام.

يؤخذ في هذه المرحلة النص المحلل معجمياً (tokenized text) ويعطى لكل مقطع وسم يدل على القسم من الكلام (POS tag) الذي ينتمي له.

* تكديس العبارات الأساسية (Base Phrase Chunking)

* تمثل هذه العملية استعادة مقدار جزئي من المعلومات النحوية لتحديد المصطلحات من جمل اللغات الطبيعية، وهي عملية تجميع الكلمات المتتالية معًا لتكوين عبارة، ولكنها لا تقدم معلومات عن كيفية اتصال العبارات معاً.

* يؤخذ في هذه المرحلة نص محلل (tokenized)، ويعطى لكل مقطع وسم يدل على نوعه من الكلام.

* قاعدة بيانات قواعد اللغة العربية: تتكون من حوالي أربعمئة قاعدة عندما تطبق على الجملة بعد استخلاص الخصائص ووضع الوسوم والحالات لكل مقطع من الجملة، وبعد تطبيق جميع القواعد إذا بقيت بعض المقاطع دون وسم ستعطى وسما افتراضيا.

بعد إجراء تقييم لنظام (بالعربي) (Bel-Arabi) وجد أن نسبة الدقة هي ٤٤ , ٩٠٪ وهي دقة مقبولة لهذه المهمة المعقدة.

International Journal of Languages, Literature and Linguistics vol. 1, no. 1, pp. 25-291, March 2015

Year of publication: 2015

Arabic between Formalization and Computation

Haytham El-Sayed

اللُغة العَربيّة بين الصياغة المنطقية الرياضية والحَوسَبة

المقدمة

يؤمن العديد من المنطقيين وفلاسفة اللُّغة بقوة بقدرة لغات البرمجة وقابلية تطبيق الطُرق الرياضية المستخدمة في نمذجتها على اللُغات الطبيعية. وعلى ضوء الفرضية القائلة بإمكانية الصياغة الرياضية للُغات الطّبيعية، فقد أمكنهم تطبيق أساليبهم المنطقية على مجموعات لُغوية طبيعيّة مُختلفة.

تمثّل هذه الورقة محاولة لتطبيق طريقة منطقية رياضية (قواعد نحو مونتاغيو **Montague**) على اللُغة العربية، كخطوة أوّلية نحو حَوسبتها. وبما أن التّمثيل الدِلالي يجب أن يكون تركيبياً أو إنشائياً على مُستوى المعالجة الدلاليّة، يمكن استخدام الصّياغة الاصطلاحية أو إضفاء الصفة الرسمية المُعتمدة على أساس (قواعد نحو مونتاغيو) كأداة مفيدة في عملية البناء الدلالي للُّغة العربية في أنظمة فهم أو استيعاب اللُّغة العربية. وفي الوقت الذي ما زالت فيه الجُهود المبذولة في سياق الصّياغة الاصطلاحية (الرسمية) النحوية–الدلالية للغة العربية محدودة جداً، تأمل هذه الدراسة في أن يمثّل هذا الأسلوب حافزا لإعادة توجيه البُحوث نحو تقنيات الترجمة الفورية الحديثة لتطوير نموذج مناسب للمُعالجة الدلالية للُغة العربية.

المُبررات ومشكلة البحث

لقد أدى النجاح في الصّياغة المنطقية الرياضية النّحوية – الدلالية للُغات المختلفة، مثل الإنجليزية واليابانية والتايلاندية، إلى إنجازات واعدة في مجال مُعالجة اللُغات الطبيعية. أما فيما

يتعلق بالعربية، فلم تتوصّل عمليات البحث لأي أدبيات منطقية ذات صلة بشأن كيفية مُعالجة اللُغة العربية بشكل منطقي رياضي. ومن الملاحظ، بالنّسبة للعربية، أن مسألة العلاقة بين المنطق واللُغة قد نوقشت كثيراً في التُراث العربي الأرسطي خلال العُصور الوُسطى. بالتالي يمكن القول بأنّ هذا العمل البحثي له جذور تاريخية تتمثّل في إعادة بناء العلاقة بين المنطق واللُغة العربية.

إضافة لما سبق، انصبّ التركيز أثناء العمل على الصّياغة المنطقيّة الرياضيّة وحَوسبة اللُغة العربية – خلال العُقود الأخيرة– على وجهة النظر الصّرفية والنحوية. وعلى الرغم من تحقق بعض النجاحات في هذا المجال، يُعتَقد بأن هناك حاجة أساسية لبذل المزيد من الجهد لتطوير نموذج مناسب للمعالجة الدلالية للُغة العربية، حيث إنه لا توجد حتى الآن نظرية اصطلاحية قائمة قادرة على تقديم صورة مُتكاملة ومُتسقة لجميع الظواهر التي تنطوي عليها المعالجة الدلالية العربيّة.

الخاتمة

حاولت هذه الورقة تقديم بعض النتائج من وجهة نظر مُؤلفها لإضفاء الطابع الاصطلاحي على جانب من قياس اللُغة العربية. ويعتقد المؤلف بأن التقدم الذي تحقّق خلال السنوات الأخيرة في مجال حوسبة اللغات الطبيعية، استناداً إلى فرضية قابلية تطبيق الأساليب االمنطقيّة الرياضيّة على اللغة العربية، يُمكن أن ينطبق أيضاً على اللغة العربية مع إجراء بعض التعديلات. وتستند الصّياغة المنطقية الرياضية للُغة العربية على (قواعد نحو مونتاغيو **Montague**) وقد استخدمت بنجاح في العديد من أنظمة مُعالجة اللغات الطبيعية لتحقيق تحليل عميق للعلاقة الدلالية–النّحوية. لكن لسوء الحظ، هناك القليل من الأعمال المعروفة لدى المجتمع اللُغوي الحاسوبي العربي للبناء الدلالي والصياغة المنطقية الرياضية للُغة العربية. لذا تأتي هذه الورقة لتُضيف حافزاً لإعادة توجيه البُحوث نحو تقنيات التركيب الدلالي الحديثة لتطوير نموذج مناسب للمعالجة الدلالية للُغة العربية.

Proceedings of the World Congress on Engineering 2009 Vol II, WCE 2009, July 1 - 3,2009 , London, U.K.

Year of publication: 2009

An Efficient Recursive Transition Network Parser for Arabic Language

Bilal M. Bataineh, Emad A. Bataineh

مُحلّل نحوي قائم على محلل شبكة التحولات المتكررة للُغة العربيّة

المقدمة

يُعدّ تحليل الجُمل العربية مُهمةً صعبة؛ وتأتي هذه الصُعوبة من عدة مصادر، أولُها أنّ الجُملة العربية طويلة ومُعقّدة، وتأتي الصُعوبات الأخرى بسبب طبيعة بنية أو تركيب الجملة العربية، فقد يكون التّركيب النّحوي لأجزاء من الجُملة مفقوداً، أو أنّها قد تتّخذ صورا مُختلفة من ترتيب الكلمات والعبارات. تهدف هذه الدراسة إلى تطوير مُحلل عربي جديد يعمل على استخلاص سِمات أو خصائص الكلمات العربية وتحليلها. وقد تمّ إعداد هذا المحلل باستخدام تقنية تحليل خوارزمية من الأعلى إلى الأسفل مع شبكة تحولات متكررة، وقد تمّت عملية تطوير المحلل من خلال خُطوتين: في الخطوة الأولى، تم إنشاء مجموعة من القواعد المستخدمة في هذا المحلّل للُغة العربية بناء على نص عربي يُدرس في الصف المدرسي الثاني عشر. أما الخُطوة الثانية فتمثّلت في تشغيل أو تطبيق المحلّل في تحليل الجملة العربية وتحديد ما إذا كانت الجُملة تتبع بنية نحويّة صحيحة أم لا. وقد تم تقييم المحلل في مقابل الجُمل الحقيقية وكانت النتائج مُرضية للغاية.

المُبررات لمُشكلة البحث

يُعدّ تحليل الجُملة العربية مُهمة صعبة، ففي عمليات مُعالجة اللُغة العربية الطبيعية، لا توجد أشكال مُحددة مُسبقاً لتحليل الجُمل، مما يجعل من عمليّة تحليل الجُملة العربية مُشكلة

مُعقّدة وغامضة من الناحية التركيبية أو النّحوية، بسبب الاستخدام المتكرر للعلاقات النّحوية، وحُروف العطف، وغيرها من التّراكيب اللُغوية.

منهجية الدراسة

استندت منهجية هذه الدراسة بشكل أساسي إلى دراسة وتحليل قواعد اللُغة العربية المطابقة للجِنس والعدد، وصياغة القواعد باستخدام قواعد اللُغة حُرة-السياق، التي تمثل القواعد المستخدمة في شبكة التحولات، وتُشكّل مُعجم الكلمات أو المفردات التي تتكون منها الجُمل، وتطبيق أو تشغيل مُحلّل شّبكة التحولات المتكررة وتقييم النظام باستخدام جُملة عربية حقيقية. واستخدمت تقنية تحليل «خوارزمية من الأعلى إلى الأسفل» مع شبكة التحولات المتكررة في تطوير المحلّل مدار البحث.

النتائج والتوصيات

الهدف الرئيسي من هذه الدراسة هو تصميم وبناء وتقييم نظام لنموذج أوّلي لتحليل الجُمل العربية، وتحديد إذا كانت هذه الجُمل صحيحة من النّاحية النّحوية أم لا. تفتقر اللُغة العربية لوجود أنظمة لتحليل الجُمل العربية، وقد أصبحت أنظمة التّحليل اللُغوي النحوي مُهمة جداً في مُعالجة اللُغات الطبيعية؛ لأنها الخُطوة الأولى في معظم تطبيقات مُعالجة اللُغات الطبيعية. أضف إلى ذلك أنه يمكن استخدام هذا النظام على نطاق واسع للأغراض التّعليمية.

وقد تم تقييم كفاءة المحلل اللغوي الذي تمّ تطويره، حيث استخدمت عينة من ٩٠ جملة في الاختبار. وأظهرت النتائج أن ٦ ,٨٥٪ من الجمل تم تحليلها بنجاح، و ٢ ,٢٪ من الجمل لم يُفلح النظام في تحليلها، و ٤ ,١٤٪ من الجمل لم يتمّ تحليلها لأسباب مختلفة، و ٤ ,٤٪ كانت فيها مشكلة مُعجمية، ونسبة ٢ ,٢٪ جمل غير صحيحة، ونسبة ٦ ,٥٪ لا يمكن التعرف عليها من اللغويين وفقا لقواعد اللغة العربية. وفي الختام كان المحلل كفؤاً وأعطى نتائج مُرضية.

TALN-2009: Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles, Senlis, France.

Year of publication: 2009

Arabic Disambiguation Using Dependency Grammar

Daoud Daoud, Mohammad Daoud

إزالة الغموض من النص العربي باستعمال نحو الاعتماد

يعرض هذا البحث طريقة لفك الغموض في النصوص العربية باستعمال مزيج من النموذج المستند للقواعد الذي يعتمد على المفهوم النحوي التقليدي واستعمال نحو الاعتماد (**Dependency Grammar**) . تساعد هذه الطريقة على الوصول إلى تحليل دقيق للجمل، منتجًا شبكة من دلالات الألفاظ باستعمال اللغة العالمية الوسيطة (**Universal Networking Language**) .

إن معالجة اللغة العربية بهدف إزالة الغموض يتم على النحو الآتي:

* يجهز المعجم بالمعلومات الأساسية حول المفردات؛ أي الخصائص النحوية (**grammatical attribute**).
* ضوابط التجاور (**Adjacency constraints**) أو عدم التوافق (**incompatibility**) مثل قاعدة أنه لا يمكن أن يتتابع حرفا جر.
* الاعتماد الصرفي (**Morphological dependencies**) وهو يصف وجوب توافق بعض المزايا، مثل أن الفعل الذي يعقب الفاعل يجب أن يتوافق معه من ناحية التذكير والتأنيث؛ أي أن الفعل معتمد على الفاعل والفاعل معتمد على الفعل.
* الاعتماد النحوي (**Syntactic dependency**).

تبدأ عملية إزالة الغموض باستعمال ظروف الاعتمادات النحوية المتجاورة، وقد استعمل برنامج نقل الجمل من اللغة العربية الطبيعية إلى اللغة العالمية الوسيطة (**Universal**

networking language (UNL))، وهي لغة ترميز لا تعتمد على لغة معينة.

لغة (EnCo) هي لغة برمجة تستند إلى القواعد، وهي متخصصة بكتابة برمجيات التحويل (converters)، وتعمل بالطريقة التالية: القيام بمسح (scanning) سلسلة النص المدخل من اليسار إلى اليمين. وفي أثناء عملية المسح يتم العثور على كل المقاطع الصوتية (morphemes) التي تبدأ بالحروف نفسها من القاموس، وتصبح مقطعا صوتيا مرشحا وفق أسبقية معينة، وذلك لبناء شبكة دلالية للجملة. أما سلاسل الحروف التي لم تمسح فيجري مسحها من البداية وفق القاعدة. إن الناتج من هذه العملية هو شبكة دلالية بلغة (UNL).

تستحضر في هذه المرحلة كل وسائل إزالة الغموض وهي المجاورة والاعتمادية الصرفية والاعتمادية الدلالية والصفات المعجمية المستخلصة من المعجم. وتلخص هذه الوسائل برموز ملحقة مع كل عقدة في البيانات المدخلة. فمثلا لجملة مثل « أخذ سامي المال كوارث شرعي» فإن كلمة (كوارث) يستنتج بتطبيق الوسائل أعلاه أنها مؤلفة من حرف الكاف (كـ) وكلمة (وارث) ولا تؤخذ (كوارث) باعتبارها كلمة واحدة.

وأثناء عملية التحويل باستعمال (converter) وصل عدد المفردات في المعجم ١٢٠،٠٠٠ كلمة. وهذا العدد يغطي معظم الكلمات العربية الشائعة. وقد أضيف لكل مفردة خواص أكثر تعقيدا، بحيث تغطي كل الجوانب النحوية والصرفية والدلالية للكلمة. وقد أخذ بعين الاعتبار عند تصميم هذه الخواص عمليتا التحويل والاسترجاع، كما أضيفت الكلمات الدلالية والسوابق واللواحق المستعملة باللغة العربية.

لقد مكننا النموذج الحاسوبي المتزامن للغة (EnCo) مع القواعد الاعتمادية من الوصول إلى الأداة المناسبة لفك ألغاز الغموض في الجملة العربية. وقد استطاع النموذج المستعمل فك غموض الكلمات بطريقة كفوءة ودقيقة بين الأسماء والأفعال وبين الأدوات والأفعال، لكنها كانت أقل كفاءة في التمييز بين الأسماء والأسماء وبين الأفعال والأفعال. وهذا متوقع نظرا لأن العلاقات الدلالية تصبح أقل حدة، مما يقلل من فاعليتها في إزالة الغموض.

IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies20-22 , Dec 2016, Khartoum, Sudan.

Year of publication: 2016

A comparative analysis of the Arabic and English verb systems using a Quranic Arabic corpus

Jawharah Alasmari, Janet C.E. Watson, Eric Atwell

تحليل مقارن لأنظمة الفعل العربية والإنجليزية باستعمال الذخيرة اللغوية للقرآن

إن الذخيرة اللغوية للقرآن الكريم هي واحدة من أهم الوسائط الحاسوبية التي أنتجت لخدمة اللغة العربية. وأهم هدف لهذا البحث هو تقديم تحليل صرفي مفصل مع تحليل تركيبي للفعل في كل من اللغتين العربية والإنكليزية. ويلقي البحث الضوء على بعض الدراسات حول استعمال الذخيرة اللغوية الجزئية للأفعال مع ترجماتها، وذلك لفهم كيف يؤشر الفعل ضمنيا على الزمن باللغة العربية وكيف يجري تحويله إلى الإنكليزية.

في البداية تمت دراسة الفعلين «قال» و «كان» في الآيات المختلفة، وبني منهما مع ترجماتهما ذخيرة لغوية ثنائية اللغة، ثم أجريت مقارنة مع مكافئاتها عند الترجمة. قامت الدراسة بعد ذلك بتأشير الأخطاء الناجمة من الترجمة، وأجريت دراسة تحليلية لهذه الجمل التي احتوت الأخطاء، ثم أجريت دراسة تكرار للتراكيب المختلفة للفعل في اللغتين. بعد ذلك جمعت ذخيرة لغوية للأفعال وسياقاتها بشكل عشوائي من القرآن. من هذه الذخيرة أمكن الإجابة على بعض الأسئلة بشكل أكثر تفصيلا، ومن هذه الأسئلة: ما هو فاعل الفعل؟ كم عدد الأفعال الدالة على المفرد أو المثنى أو الجمع؟ ما هي الأفعال الدالة على المذكر أو المؤنث؟ هل الترجمة الإنكليزية تستعمل نفس الصيغة؟ هل الصيغة عند تحليل السياق تعطي نفس صيغة الفعل؟

بالطبع بعض هذه الأسئلة لم يكن جوابها بسيطا: نعم أو لا.

لقد اختيرت ٦٠ جملة ضمن شروط معينة تتضمن الفعل «يقول» لبدء عملية دراسة

الاختلافات بين اللغتين، وذلك لمقارنة الزمن والتذكير والتأنيث والإفراد والتثنية والجمع. هذا بالإضافة لمعرفة أفضل ترجمة مناسبة للفعل باستخدام عدد من ترجمات القرآن الكريم.

الخلاصة

عند إلقاء نظرة شاملة يمكن استنتاج أن الذخيرة اللغوية للقرآن الكريم هي واحدة من أهم الأدوات الحاسوبية التي أنتجت في خدمة اللغة العربية، فهي توفر للمتعلمين ما يحتاجون إليه في مجال اللغة واللغويات والدراسات الحاسوبية. كما تمهد الطريق للباحثين لدراسة الهياكل الصرفية والنحوية من خلال دراسات الحوسبة العميقة للقرآن الكريم. على وجه الخصوص، نوقشت الاختلافات وأوجه التشابه بين النظم العربية والإنجليزية للأفعال التي تساعد على تقديم بعض التفاصيل التي سيتم استخدامها لتحسين الترجمة الآلية من اللغة العربية إلى اللغة الإنجليزية.

The 13th International Arab Conference on Information Technology ACIT'2012
Dec.10-13

Year of publication: 2012

Parse Trees of Arabic Sentences Using the Natural Language Toolkit

Maad Shatnawi, Boumediene Belkhouche

أشجار التحليل للجمل العربية باستخدام حزمة برمجيات اللغة الطبيعية

هناك تطبيقات عديدة للتحليل النحوي الكامل (Parsing) وإزالة الالتباس الدلالي، والترجمة الآلية وتوليد الكلام. وتعد قواعد اللغة العربية التقليدية واحدة من الأصول التاريخية لقواعد نحو الاعتماد الحديث، وهناك العديد من المحللات النحوية (parsers) للغة العربية الحديثة، وكلها تحتاج إلى بنك للأشجار النحوية (Treebank) .

ويواجه التحليل النحوي للغة العربية العديد من التحديات بما في ذلك السمات الفريدة للغة، والدرجة العالية من الالتباس في نظام الكتابة للغة العربية، والصرف؛ إذ إن عملية تشكيل الكلمات معقدة للغاية من الجذور والأوزان، وطول الجمل، والترتيب الحر للكلمات في الجملة إضافة إلى وجود الضمائر المستترة. ونتيجة لذلك، يمكن أن يقترح المحلل البديل معاني بديلة لجملة معينة. ومن أهم تحديات التحليل النحوي للغة مثل اللغة العربية هو التجزئة الصرفية إلى أجزاء؛ إذ تشكل هذه الأجزاء وحدات فردية في التحليل النحوي وعقد أوراق منفصلة في شجرة بناء الجملة .(Syntax tree)

اللغة العربية الحديثة (Modern Standard Arabic) هي اللغة العربية المعيارية المستخدمة حالياً في الشرق الأوسط وشمال إفريقيا، وإحدى اللغات الرسمية الست في الأمم المتحدة. أما اللغة العربية القرآنية فهي اللغة العربية المستخدمة في القرآن، وهي الأساس المباشر للغة العربية الحديثة. يواجه بناء تحليل نحوي مشجر للقرآن الكريم مجموعة إضافية من التحديات مقارنة

باللغة العربية الحديثة. ويرجع ذلك إلى حقيقة أن القرآن الكريم أكثر تنوعاً من اللغة العربية الحديثة في جوانب كثيرة؛ منها الجانب الإملائي (Orthography) والتهجئة (Spelling) والتصريف (Inflection) . ولا يزال التحليل الحاسوبي للقرآن الكريم مجالاً مثيراً للاهتمام لكنه لا يزال غير مستكشف بسبب عدم بذل جهود كافية وإن وجدت فهي فردية غالبا.

في هذه الورقة البحثية، قدم الباحثان نهجاً لتوليد أشجار تحليل من الجمل العربية بشكل عام والجمل القرآنية على وجه الخصوص، وذلك باستخدام مجموعة أدوات اللغة الطبيعية. فتألفت العملية المحددة من بناء معجم، وقواعد خالية من السياق النحوي، واستدعاء مجموعة أدوات اللغة الطبيعية المكررة، ويمكن النظر إلى أشجار التحليل كمكونات بنك الأشجار النحوية (Treebank)، والطريقة المقترحة تدعم بناء المكنز اللغوي للقرآن الكريم. وقد نجح النموذج المقترح في تحليل السور القرآنية بالإضافة إلى جمل اللغة العربية الحديثة، ويدّعي الباحثان أن توفر هذا المحلل سيكون مفيداً في مختلف تطبيقات معالجة اللغة الطبيعية مثل الترجمة الآلية، واسترجاع المعلومات، كما يمكن توسيع نطاق عمل الباحثين في عدة اتجاهات مثل دمج المحلل مع محلل صرفي من شأنه أن يزيد من أتمتة العملية.

International Journal on Information and Communication Technologies, Vol. 3, No. 3, June 2010

Year of publication: 2010

Rule-based Approach in Arabic Natural Language Processing

Khaled Shaalan

النهج القائم على القواعد في معالجة اللغة العربية الطبيعية

استخدمت المنهجيات القائمة على القواعد (Rule-based Approach) بشكل ناجح في تطوير العديد من نظم معالجة اللغات الطبيعية (Natural Language Processing) ، ويتطلب العمل في هذا النهج عادة معرفة لغوية قوية، لا تحتاج إلى ذخيرة لغوية كبيرة لإتمام مهمتها، حيث إن توفير كمية كبيرة من الذخيرة اللغوية ليس بالأمر السهل نظرا لقلتها وصعوبة الحصول عليها. وقد دفع هذا الكثير من الباحثين إلى اعتماد هذا النهج في تطوير انظمتهم المتعلقة بمعالجة اللغة الطبيعية. ومن الانتقادات التي وجهت إلى هذا النهج أنه موضوع تقليدي وتمت دراسته بشكل واسع وخاصة في اللغات الأوروبية. ولكن اللغة العربية لم تحظ بالكثير من البحث في هذا المجال ولم يتسن لها الاستفادة الكاملة من ميزاته بشكل كبير في السابق، فكان لابد من بحثه ودراسته في اللغة العربية.

في هذه الورقة البحثية، عرض الباحث الجهود الناجحة التي أنجزت لمختلف مهام معالجة اللغة العربية باعتماد النهج القائم على القواعد. ووضح الباحث تجاربهم المتنوعة في تطوير أنظمة وأدوات ناجحة تعتمد على قواعد معالجة اللغة العربية. ويشتمل هذا النهج على عدة أدوات أساسية قائمة على استخدام محلل نحوي (syntactic analyzer) ومحلل صرفي (morphological Analyzer) ومترجم آلي (machine translators) وأداة التعرف على أسماء الأشياء (named entity recognizers) إضافة إلى الأنظمة المحوسبة لتعلم اللغة. كما

أن هناك بعض المهام الخاصة التي يقوم بها هذا النهج مثل تحويل اللهجات العربية المختلفة إلى اللغة العربية الفصحى، مثل اللهجة المصرية أو السعودية وغيرها. ويفرض هذا النهج قيودا لغوية مقنعة، ويسمح باستخدام قواعد الاستدلال (heuristics)فمثلا الفعل لا يجوز أن يسبقه حرف جر، ويعتمد على قواعد مكتسبة من المختصين اللغويين وليس بناء على تدريب تلقائي من البيانات. كما أن هذا النظام يقوم على دمج المعرفة اللغوية مع قواعد المجال التي توفر نتائج عالية الثقة والدقة، فهذه القواعد تعمل على تكوين الجمل العربية وتحليل المدخلات غير السليمة.

وبغض النظر عن ندرة البيانات، فإنه بالمقارنة مع المناهج القائمة على تعلم الآلة (machine learning approaches) التي تواجه بعض العوائق المتعلقة بمعالجة اللغات الطبيعية، فإن النهج المقترح يستطيع التغلب عليها وحلها، فعلى سبيل المثال في البرامج التعليمية على الإنترنت يصعب على النهج القائم على تعليم الآلة التمييز بين المدخلات الصحيحة والمدخلات غير الصحيحة، أما بالنسبة للنهج القائم على القواعد فيستطيع تزويد المتعلم بتحليل واف للإجابة باستخدام المعرفة اللغوية مما يزوده بتغذية راجعة لمساعدته في فهم أفضل. وقد أوضحت التجارب والدراسات التي تم إجراؤها من القائمين على هذا المشروع أن هذا النهج يتطور بشكل سريع خاصة في غياب المعرفة اللغوية وبعض المشكلات المتعلقة بتكييف الأدوات المستخدمة في لغات أخرى عند تطبيقها على اللغة العربية، ويمكن التغلب عليها مستقبلاً.

٣-٣ أبحاث الموارد

تتوزع هذه الأبحاث إلى خمسة مواضيع فرعية هي: المعاجم الآلية، والمكنز، والمدونات الموسومة، والأنطلوجيا، وشبكات الكلمات.

٣-٣-١ أبحاث المكانز والذخائر اللغوية

وتضم ٨ أبحاث بينها بحث من نوع الأبحاث المسحية عنوانه دراسة مَسحية نقديّة حول الذخائر اللغوية العَربية المُتاحة مجاناً وبحث واحد نوع أ عنوانه: تكوين ذخيرة لغوية عربية للأطْفال.

أما الأبحاث نوع ب فعددها ستة أبحاث هي: نحو مقياس لجودة الذخيرة اللغوية العربية، و الذخيرة اللغوية لدارسي اللغة العربية «(ALC) v2» – ذخيرة جديدة مكتوبة ومنطوقة لدارسي اللغة العربية، وبناء الذخيرة العالمية للغة العربية «ICA» : تقدم مرحلة التجميع، والتنبؤ بغموض التنسيق في اللغة العربية المستند إلى ذخيرة لغوية، و مُعالجة ذخائر النُصوص العربية الكبيرة: تحليل أولي مع النّتائج.

Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Reykjavik, Iceland, 26-31 May 2014

Year of publication: 2014

Critical Survey of the Freely Available Arabic Corpora

Wajdi Zaghouani

دراسة مَسحية نقديّة حول الذخائر اللغوية العَربية المُتاحة مجاناً

المقدمة

يُعدّ توفر الذخائر اللغوية (corpra) عاملاً رئيسياً في بناء تطبيقات مُعالجة اللُغة الطّبيعية، لكن تكاليف الحُصول عليها يُمكن أن تحُول دون مضيّ بعض الباحثين قُدماً في مَساعيهم لتوفيرها. وتُعدّ سُهولة الوصول إلى الذخائر اللغوية المُتاحة بشكل مجّاني حاجةً مُلحةً لدى مُجتمعات الأبحاث التي تعمل على معالجة اللغات الطبيعية، خاصة بالنّسبة للُغة مثل اللُغة العربية. لكن ليس من السّهل – في الوقت الراهن – الوصول إلى قائمة شاملة ومُحدّثة من الذخائر اللغوية العربية المتاحة مجاناً.

كما أن مُعظم الموارد التي يُوفرها مُزودو بيانات اللّغات إما أن تكون مُقابل رسوم أو مُخصّصة حصرياً للمُشتركين، مثل تلك الذخائر اللغوية التي قام بتوفيرها اتحاد البيانات اللُغوية (LDC) أو وكالة توزيع الموارد واللُغات (ELDA). وأظهر استعلام بسيط عن اللغة العربية المتاحة في فهرس اتحاد البيانات اللغوية توفر ١١٦ ذخيرة لغوية من أنواع مختلفة (النّصية، والكلامية (الخطابية)، والتقييمية، الخ ...). ولدى إجراء استعلام آخر مُماثل، أظهر مُحرّك البحث (ELRA Corpora) توافر ٨٠ ذخيرة اللغوية. وما تزال اللُغةُ العربية لُغةً فقيرةً نسبياً بالموارد بالمقارنة مع لغات أُخرى مثل اللُغة الإنجليزية، وسوف يُساهم الوصول إلى الذخائر اللغوية المُتاحة مَجاناً

بالتّأكيد – في تحسين تقنيات المعالجة اللُغوية العربية الموجودة حالياً. وتعرض هذه الورقة نتائج دراسة مسحيّة عبر الشّابكة حول الذخائر اللغوية للُغة العربية المُتاحة مجاناً.

المبحث الأول: الوضع الراهن للذخائر اللغوية العربية المُتاحة بشكل مجاني

قبل البدء في هذا المسح، أجريت عدة استعلامات عبر الشّابكة لتحديد موقع أي من الذخائر اللغوية العربية المُتاحة بشكل مجاني أو أي مُستودع وثائق يضم قائمة بالموارد التي يُمكن الوصول اليها بسهولة. وجدت أن المعلومات مُتناثرة بين مُختلف المواقع الشّخصية والمَجموعات البحثية، التي غالباً ما تكون غير كاملة أو عفا عليها الزّمن.

في مطلع العام ٢٠١٠، أنشأ مُحرّك البحث (ELRA Corpora) خارطة التقييم والموارد اللُغوية (LRE) وهي عبارة عن قاعدة بيانات إلكترونية (عبر الشّابكة) للموارد اللُغوية. والهدف من هذه الخارطة هو رصد إنشاء موارد اللُغة واستخدامها. وجُمعت المعلومات المتعلقة بها أثناء عملية تقديم الطّلب لـمُؤتمر (LREC) ولغيره من المُؤتمرات. لقد قُمنا بإجراء استعلام عن قائمة الذخائر اللغوية العربية المتاحة مجاناً، وعثرنا على عدد محدود، ولم يتوافر لنا أي عنوان لرابط المواقع على الشبكة (URL) يربطُنا بالمستخدم أو بتفاصيل المشروع أينما وُجدت. ويوجد لدى جمعية اللُغويات الحاسوبية (ACL) صفحة «ويكي» تسرد الموارد المتاحة حسب اللُغة، أما الصّفحة العربية فكانت تضم فقط خمس ذخائر، منها أربع ذخائر مجانية، وذخيرة واحدة خاصة (غير مجّانية).

وتحتفظ الشّبكة الأوروبية للتميّز في تقنيات اللُغة البَشرية (ELSNET) بقائمة من المؤشرات التي تدُل على اللُغة العربية وغيرها من البرامج المتعلقة باللغات الساميّة ومواقع الكلام، في حين اشتمل قسم الموارد العربية على ٢٣ إدخالاً مُعظمها أُنشئت منذ ما يزيد عن ١٢ عاماً.

من جهة أخرى، قام مشروع تقنيات الكلام واللُغة العربية المتوسطية (MEDAR) بإجراء مسح في عامي ٢٠٠٩ و ٢٠١٠ لإعداد قائمة بالمؤسسات والخبراء العاملين على تطوير موارد اللُغة العربية، بالإضافة إلى الأنشطة والمشاريع التي يجري تنفيذها وما يتصل بها من أدوات. ومن ثمّ جُمعت النتائج ووضعت على قاعدة البيانات التي يمكن الوصول إليها عبر الشّابكة.

ومرة أخرى، على الرغم من الجهود الكبيرة التي بُذلت، توقّف تحديث القائمة، كما أنها تفتقر إلى المعلومات الضّرورية لتحديد موقع البيانات مثل صفحة التنزيل أو وصف المشروع وعنوان الرابط على الشابكة (URL) وأخيراً، كان هناك جهود شخصية أخرى مُتميّزة لإدراج بعض موارد اللُغة العربية مثل «مُستودع سيبويه لمعالجة اللُغة العربية»، وموقع «صفحة السُليطي لمتون اللُغة العربية» وموقع «صفحة الغامدي للروابط العربية». وتُعتبر جُهودنا الموصوفة في هذا المشروع تكملة لما هو موجود فعلياً مع التركيز على الموارد المتاحة مجاناً، وكيفية الحُصول عليها.

المبحث الثاني: المسح

لأجل البِدْء في جَمْع قائمة بالذخائر اللغوية العربية، أُجريَ هذا المسح على الشّابكة بمشاركة مُختلف القوائِم ذات الصّلة بمعالجة اللغات الطبيعية مثل Corpora و ArabicList. وكان من المفترض استكمال هذه الدراسة المسحيّة على الشّابكة خلال ٥–١٠ دقائق لتشجيع المشاركين، واشتملت على بعض الأسئلة الأساسية جداً مثل معلومات المزود (providerinformation)، ونوع وحَجم والغَرض من الذخيرة، ورابط التحميل (download link)، والمنشورات ذات الصلة، والتنوّع العربي، وحالة الإنتاج (production status)، والتّأكيد على أنّ الذخيرة مُتاحة بشكل مجّاني تماماً لأغراض البحث. وشارك في المسح الإلكتروني ٢٠ مشاركاً أشاروا إلى حصيلة موارد بلغت ٢٦ موردًا. وبمُجرد تجميع نتائج المسح، أُضيف ٤٠ موردًا آخر من الموارد العربية المتاحة مجاناً والمأخوذة من موارد مُختلفة عن الشابكة، وستتم مناقشتها في المبحث التالي. وأخيراً حاولت الدراسة من خلال هذا المسح تحديد موقع أي مادة منشورة ذات صِلة بالموضوع، وسيقدم وصف موجز لمجموعة مُختارة مكوّنة من ٦٦ موردًا من الموارد المجّانية التي عُثر عليها في هذا المسح.

المبحث الثالث: الموارد المتاحة

استعرضت الدراسة في هذا القسم نتائج مسح الذخائر اللغوية العربية المتاحة بشكل مجاني مع التركيز على أهم عمل لكل فئة من الفئات التالية:

* الذخائر اللغوية النصية الخام (Raw Text Corpora).
* الذخائر اللغوية المَشروحة (Annotated Corpora).
* المُعاجم (Lexicon).
* الذخائر اللغوية الكَلامية (الشّفهية) (Speech Corpora).
* ذخائر تمييز خط اليد (Handwriting Recognition Corpora).

أنواع أخرى مُتفرقة من الذخائر اللغوية.

استعرض هذا المبحث بعض المعلومات الأساسية لكل فئة من هذه الفئات الأربع، اشتملت على اسم المؤلّف أو المجموعة البَحثيّة، واسم الذخيرة ، وحجم الذخيرة حسب عدد الكلمات أو عدد الملفات. وفي حالة وجود مادة منشورة مُرتبطة بالهيئة، يُستشهد بها جزءاً يتبع اسم المؤلف الوارد في الجدول، وإلاّ يُذكر عنوان رابط الوصول/ التّنزيل على الشابكة (URL) في الحاشية السُفلية بعد اسم المتن في كل جدول. ومن ثمّ فُرزت الموارد في الجداول وفقاً لحجمها حسب الأهمية (من الأكثر أهمية إلى أقلها أهمية).

١- ذخائر لغوية للنُصوص الخَام: في هذا القسم جرى الاستشهاد بـ ٢٣ ذخيرة للنُصوص الخام المتاحة مجاناً، أي أنها لا تتضمن أيا من تلك النُصوص التي لا تتضمن أي نوع من الشّرح، وتقتصر على الملفات النّصية نفسها. وتنقسم هذه الذخائر اللغوية إلى أربع فئات كما هو مبين أدناه:

- الذخائر اللغوية أُحادية اللغة: وتشتمل على ١١ ذخيرة أحادية اللغة مُتاحة كلُها للتنزيل بشكل مجاني.
- الذخائر اللغوية ثُنائية اللغة: من بين الذخائر اللغوية الواردة في هذا المسح يمكننا التركيز على ذخيرة الأمم المتحدة باعتبارها الأهم والأكثر انتشاراً ذخيرةً متاحة مجاناً من بين الذخائر اللغوية لهذه الفئة. هُناك أيضاً الذخيرة اللغوية (nadeeM) وتشتمل على مليون كلمة عربية / إنجليزية مقترنة ضمن جملة واحدة، وتعدّ من ضمن الموارد

القيمة جداً. إضافةً إلى ذخيرتي الحديث النبوي والقرآن الكريم المتاحتين باللغتين العربية / الإنجليزية والمتوفرتين على أداة الترجمة المصرية وهي الموارد الأقل شهرة التي يمكن استخدامها في أي عمل يتعلق بالمجال الديني.

- الذخائر اللغوية باللهجات المحلية: اقتصرت هذه الدراسة المسحية على ذِكر ذخيرتين لغويتين فقط من الذخائر اللغوية باللّهجات المحلية، وذلك – حسب الدراسة – لأن مُعالجة اللّهجات العربية تُعتبر مُهمة حديثة نسبياً، وهناك حاجة فعلية لمثل هذه الموارد. وتجدر الإشارة إلى ذخيرة تحوي نحو ٠٠٠٣ كلمة من اللّهجة التُونسية، في حين اشتملت الذخيرة الثانية على نحو مليوني كلمة غريبة جُمعت من ٥٥ ألف صفحة على الشابكة وتغطي أربع لهجات عربية رئيسية: (الخليج العربي و بلاد الشام وشمال أفريقيا ومصر).
- الذخائر اللغوية المنشورة على الشّابكة: في هذه الفئة وُضعت بعض الذخائر اللغوية المتاحة حَصْرياً على الشّابكة من خلال واجهة استعلام، بحيث لا تكون هناك بيانات متاحة للتحميل يمكن أن تُسبب تشويشاً لبعض الدراسات البحثية، لكن مثل هذه الذخائر اللغوية المتاحة على الشابكة يُمكن أن تكون ذات قيمة كبيرة.

٢- الذخائر اللغوية المَشروحة: الذخائر اللغوية المشروحة مُفيدة جداً لإنشاء أنظمة وأدوات قائمة على الخوارزميات الخاضعة للإشراف، كما أن توفر تلك الموارد بشكل مجاني من شأنه أن يُساعد الباحثين الشّباب على التدرُب وبناء أنظمة بأقل تكلفة مُمكنة. واستعرضت الدراسة ضمن هذا القسم أنواع الذخائر اللغوية التالية: الكيانات (الهيئات) المسمّاة، أقسام الكلام (POS)، والذخائر اللغوية المؤشرة نحويا (annotated corpora syntactically) والمؤشرة دلاليا (semantically annotated).

٣- المعجم: في هذا الجزء ناقشت الدراسة قواعد البيانات المعجميّة (lexical data base) وقوائم الكلمات، حيث كان مُعظم هذه الموارد مُتاحاً للتّنزيل بشكل مجاني، وحسب نتائج

المسح: كان بعض تلك الموارد جُزءاً من أنظمة وأدوات، وبما أن تلك الأدوات مفتوحة، يُمكن الإفادة من تلك المعاجم للأغراض البَحثيّة.

٤- الذخائر اللغوية الكَلامية (الشّفهية): اشتملت على التّسجيلات الصّوتية، والبيانات اللّفظية المنسوخة أو المكتوبة.

٥- الذخائر اللغوية لتمييز خط اليد: الوثائق الممسوحة ضوئياً (handwriting recognition corpora) والوثائق المشروحة أو المُفسّرة.

٦- الذخائر اللغوية المحدّدة للخطأ: يمكن الإفادة منها في دراسات الأخطاء المعتمدة على الذخائر اللغوية، وكذلك في إنشاء أدوات تصحيح لغوي تلقائية.

٧- أنواع أخرى متفرقة من الذخائر اللغوية: أضيفت إلى نتائج المسح ٧ ذخائر لغوية أخرى مُفيدة للمهام ذات الصّلة بمعالجة اللغة العربية مُتعددة-الاتجاهات، مثل الأسئلة / الإجابات، ذخائر المقارنات، وذخائر الكشف عن القَرصنة الأدبيّة والمُلخّصات.

الخاتمة

عرضت هذه الورقة نتائج دراسة مَسْحية أُجريت حديثاً لتحديد قائمة الموارد والذخائر اللغوية المتاحة باللُغة العربية، فهدفت الدّراسة إلى تعزيز استخدام الذخائر اللغوية المجانيّة، وخاصة من الأفراد الذين يفتقرون إلى التّمويل ولا يتمكنون من تحمّل تكاليف الاشتراك أو العُضوية أو الرُسوم الباهظة للحُصول على إحدى الذخائر اللغوية من مراكز البيانات اللّغوية. وأظهرت النتائج الأولية للمسح وجود قائمة أوّلية مكوّنةً من ٦٦ مورداً تغطي الفئات الرئيسية من أنواع الذخائر اللغوية المختلفة. وقد عُرضت نتائج مسح الفئات المختلفة المدروسة، كما جرى توفير روابط مُباشرة للحُصول على تلك البيانات قدر الإمكان. وأظهرت النتائج أن العديد من الموارد العربية المتاحة بشكل مجّاني ليست مُتاحة دائماً، لذا يصعُب على المستخدمين المحتَملين العُثور عليها.

ويأمل القائمون على هذه الدراسة أن تكون هذه المحاولة المبدئية لتحديد مواقع أو عناوين الذخائر اللغوية العربية المتاحة مجاناً مُفيدة للمُجتمعات البحثية. وقد أُدرجت قائمة الذخائر اللغوية المعروضة في هذا البحث ضمن صفحة واحدة على الشّابكة، لسهولة الوصول إليها. وهناك خُطة مُستقبلية لجعل القائمة مُتاحة ضمن قاعدة بيانات على الشابكة، وأوصت الدراسة باستمرار جهود البحث عن ذخائر لغوية مجانية أخرى لإغناء مستودعها من الموارد العربية.

Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),May,23-28, 2016,Portorož, Slovenia.

Year of publication: 2016

Compilation of an Arabic Children's Corpus

Latifa Al-Sulaiti, Noorhan Abbas, Claire Brierley, Eric Atwell, Ayman Alghamdi

تكوين ذخيرة لغوية عربية للأطْفال

مقدمة

فكرة هذه الورقة البحثية مستوحاة من الذخيرة اللغوية لأكسفورد للأطفال (Oxford Children's Corpus (OCC))، حيث قام الباحثون بتطوير نموذج أولي لذخيرة لغوية من النُصوص العربية المكتوبة و / أو المختارة للأطفال، فجمعت وثائق للطفل العربي، بلغ عددها ٢٩٥٠ وثيقة وما يقرب من ٢ مليون كلمة، يدوياً من الشّابكة خلال مشروع مدته ٣ أشهر. وهي نُصوص ذات جَودة عالية، وتحتوي على مجموعة مُنوّعة من حكايات من موارد مختلفة، بما في ذلك حكايات تقليدية (كلاسيكية) مثل حكايات ألف ليلة وليلة العربية، وحكايات عن شخصيات خيالية شعبية مثل جُحا. ونتوقع أن هذه الإصدارات وغيرها من الإصدارات اللاّحقة من الذخائر اللغوية سوف تقود لدِراسات مُهمّة في تصنيف النُصوص واستخدامات اللُغة وأيديولوجية نُصوص الأطفال.

عندما نشرت مجلة أوكسفورد الذخيرة اللغوية للأطفال في عام ٢٠١٢، عدّت أولَ عمل من نوعه، والتعريف المفضل لأدب الأطفال الذي تمّ تبنيه لتلك الذخيرة اللغوية هو أنها: مادة مكتوبة خاصة بالأطفال، و/ أو مختارة للأطفال من الآباء والمعلمين والناشرين. وتضمّن الإصدار الأول منها حوالي ٣٠ مليون رمز أو إشارة من النصوص الإنكليزية للأطفال الذين تتراوح أعمارهم ما بين ٥-١٤ سنة، وقد نُظمت ضِمن أربع فئات أساسية (الخياليّة؛

وغير الخيالية؛ والكتابة الخاصة بالأطفال؛ وغير المصنفة)، مع التركيز على مواد القرن الحادي والعشرين. ويبدو أن الذخيرة اللغوية للأطفال على النحو السابق كانت قليلة أو نادرة. وكما هو الحال في مجموعة أوكسفورد للغة الإنجليزية، تعد الذخيرة اللغوية العربية للأطفال مدار بحث هذه الورقة هي الأولى من نوعها باللغة العربية.

في هذه الورقة البحثية تمّت مناقشة المحاور الرئيسة الآتية:

أولاً: مُبرّرات جمع ذخيرة لغوية خاصة بأدب الأطفال

ابتداءً جُمعت الذخيرة اللغوية من أوكسفورد للأطفال لتكون أساساً لأغراض المعجم، وللإعلان عن قواميس للأطفال لدى مطبعة جامعة أكسفورد. وكانت هناك حاجة إلى ذخيرة لغوية كبير للمُساعدة في تنقيح قوائم الكلمات الرّئيسية، وتحديد المُرادفات والمعاني والأمثلة الموجودة بشكل طبيعي للمفردات المستهدفة المستخدمة في السياق. وكان مدار النقاش حول عدم اقتصار قواميس الأطفال على تبسيط لغة قواميس الكبار فقط، لكن أن تُبنى أيضاً على ذخائر لغوية خاصة للأطفال التي تعكس بدورها اللغة والمحتوى الذي يتعرض له الأطفال بشكل صحيح. ومن الواضح أن نصوص الأطفال تكون عادةً ذات صِبغة تربوية وذات معان تعليمية، وهذا ينطبق أيضاً على أدب الأطفال العربي. وأحد أكثر الأنواع شعبية إلى جانب النُصوص الدّينية (مثل قصص حياة النبي وحياة الأنبياء الآخرين مثل موسى وعيسى) الخيال التاريخي الذي يتحدث عن الأبطال المسلمين، وهذا ما تُعززه دراسة حول مجلات الأطفال العربية.

ما زال مشروع تأليف أو تصنيف ذخيرة لغوية عربي للأطفال (ومشروح لاحقاً) مُستمراً. ومع ذلك، فإن النُسخة الأولى من الذخيرة لغوية، التي تشتمل على رموز على ١٨٧٧٦١٥ كلمة رمزية، تم استخراجها حَصرياً من الشّابكة، وهي تُحاول بالفعل تصنيف مجموعات أوسع من أنواع بيانات لغة الأطفال أكثر من أي مجموعات بيانات سابقة.

ثانياً: تحديد موارد الذخيرة اللغوية

أوّل مرحلة لتصميم الذخيرة اللغوية هي البَت في موارد النُصوص. وبما أن الشّابكة غنيّة

بالنُصوص العَربية التي يمكن الوصول إليها بسهولة، فقد كانت المورد الرئيسي لجمع نُصوص الذخيرة اللغوية. وقد تمكنت هذه الدراسة من حصر حوالي ٧٠ موقعاً لجمع النُصوص منها، وقد استُبعدت المواقع التي تَحتوي بشكل أساسي على ملفات الصّوت والفيديو. كما استُبعدت ملفات الـ pdf الممسوحة ضوئياً (OCR) نظراً لعدم وجود نظام دقيق للتعرُف على الحُروف العَربية الضّوئية. صُنّفت قائمة المواقع المحصُورة باستخدام سِياق مُحدّد، حيث أظهرت النتائج أن مُعظم المواقع التي عُثر عليها هي مواقع لمجلات الأطفال على الشّابكة ومنتديات تضم أقساما صغيرة لقصص الأطفال ومُدوناتهم التي أُنشئت من أشخاص وضعوا عليها كتاباتِهم أو جمعوا فيها قصصاً كتبها آخرون. وبعد تحديد المواقع المناسبة، قام كل باحث شارك في جمع النصوص بالتأكد من الاحتفاظ بسجلّ يحتوي على عنوان كلّ نص ومصدره على ملف إكسل.

ثالثاً: جَمْع الذخيرة اللغوية

كانت الخطّة المبدئية لجمع الذخيرة اللغوية قائمة على أساس استخدام برنامج WebBootCat ، وقد وفّر ذلك بديلين لجَمع الذخيرة اللغوية:

(١) عن طريق جذور المفردات الأصلية المستخدمة في الاستعلام من خلال مُحرك البحث غوغل، حيث شكّلت النّتائج اللاّحقة ذخيرة لغوية أولية؛ (٢) من خلال تحميل عناوين الروابط على الشابكة (URL) ، لكن لوحظ في بعض الأحيان أنه عند تنزيل النّص من الشّابكة، يكون النّص ناقصاً؛ وعند مقارنته بالوثيقة الأصلية إما أن تكون بداية النص أو نهايته مفقودة. وبالتالي وُجد أن الأداة فشلت في بعض الأحيان في معالجة مواقع الشّابكة التي يحتمل أن تكون ذات صلة بالبحث. ولذلك تقرر في نهاية المطاف أن أفضل طريقة لجمع الذخائر اللغوية للأطفال هي القيام بذلك يدوياً. وبهذه الطريقة، ضُمن أن تكون النُصوص كاملة وخالية من أية مواد غير مرغوب فيها.

رابعاً: تَصنيف الأنواع

تُقدم هذه الذخيرة اللغوية ملامح جديدة ومتنوعة لأدب الأطفال العرب في القرن الحادي

والعشرين، وقد تضمّن الإصدار الأول من هذه الذخيرة اللغوية فئتين من النصوص: الخيالي أو غير الخيالي، ولكل فِئة قائمة من الأنواع الخاصّة بها. غير أن تصنيف النُصوص لفئاتها الأساسية لا يحول دون التّصنيف الفرعي لها. على سبيل المثال، كيف يُمكن تصنيف سلسلة هاري بوتر؟ وهي رغم تصويرها للحياة العادية كما نعرفها (المدرسة والمعلمين والعطلات والمنزل، الخ)، إلا أنّها تَستحضر أيضاً عالم الخَيال والقُوى السحرية. فهي ليست مُجرد كُتب للأطفال، فالأمّهات والآباء (والناس المسافرون على متن القطار) يُمكنهم أيضاً قراءتها! لذلك تضمنت الخطة أن تخضعَ الذخيرة اللغوية العربية مدار البحث لمزيد من التّنقيح في الإصدارات اللاّحقة كونه مشروعا مُستمرا.

خامساً: عَرْض أو تقديم الذخيرة اللغوية (التّنسيق والتّخزين)

جرى تحميل الإصدار الأوّل للذخيرة اللغوية العَربية للأطفال على أداة الاستعلام **SketchEngine** استعداداً لمزيد من التحقُق والتَحليل، وذلك وفقاً للخُطوات الآتية:

* جُمعت القِصص من مواقع مُختلفة على الشابكة، وأزيلت جميع التّنسيقات النّصية منها.
* لاحقاً، أضيفت ترويسة لكل قِصة، بحيث تحتوي على الحقول التالية: عنوان رابط (**URL**) للقصة، وعنوان القصة، والمؤلف، والنّوع، واللهجة.
* ثم يقوم محرك الرسم التخطيطي (**SketchEngine**) بتكوين ملف ضبط لكل ذخيرة لغوية يتم تحميلها، بحيث يقع ملف الضبط في دليل السّجل مع اسم الملف الذي يُوافق اسم الذخيرة اللغوية على النظام. يحتوي الملف على المعلومات الأساسية مثل اللغة والتشفير وكذلك تعريفات الصّفات التي تتعلق بالعَلامات الفَوقية المضافة لملفات القصة.
* بعد ذلك وُضعت علامة عند بداية كل نص ونهايته لتسهيل استخدام أداة (**Onion**) التي يُوفرها مُحرك (**SketchEngine**) لإزالة أي تكرار في النُصوص عند تأليف الذخيرة اللغوية.

سادساً: الاستنتاجات والتوصيات

استعرضت هذه الدراسة تصنيف أوّل مجموعة من نصوص الأطفال العربية، التي تُعرّف بأنها نُصوص مكتوبة أو مختارة خِصيصًا للأطفال. وتعد البيانات التي جُمعت يدوياً ذات جَودة عالية وتغطي قائمة واسعة من الأنواع. وقد مثّل جمع الذخيرة اللغوية هذه تحدياً كبيراً بالنّسبة للمؤلفين، حيث لم تكن مسبقًا توجد طريقة واضحة لتوصيف نصوص الأطفال، ولم تكن الأساليب الآليّة (المستخدمة في نظام (**WebBootCat**) التي تستخدم -على سبيل المثال- المفردات الأصليّة لتحديد عناوين **URL** وتحميلها على الشّابكة(ناجِحة دائماً في تصفية المواد غير المناسبة و / أو غير المرغوب فيها. لقد اختار المشروع منصة (**SketchEngine**) لأغراض تنسيق الذخيرة اللغوية وتخزينها، وهي إحدى المنصات الرائدة في السوق التي تتيح تصدير البيانات الأساسية وفقاً لعدة أشكال أو خيارات مُختلفة.

وقد أوصت الدراسة بتطوير هذه الذخيرة اللغوية وتنقيحها أيضاً في سياق مشروع كبير قائم على التطبيقات والبرمجيات الحاسوبية، بحيث يتضمن الشّرح التلقائي والتحليل، على سبيل المثال، بمقارنته مع اللغة العربية للكبار، من خلال استخلاص المفردات وتسلسلات الصيغ (**formulaic sequences**) في النصوص المكتوبة للأطفال، ومقارنتها مع المفردات وتسلسلات الصيغ في النصوص العربية المكتوبة للبالغين.

وختاماً، فإن المؤلفين يعتقدون أنه مع مزيد من التطوير على هذه الذخيرة اللغوية الخاصة بالطفل العربي، سوف يشكل ذلك رافداً ممتازا للبيانات المطلوبة لمزيد من الأبحاث في مجال تعليم اللغة العربية وعلومها، وتطوير كُتب القراءة للأطفال العرب.

Proc. Int. Colloquium on Arabic Language Processing, CITALA-2007, Rabat, Morroco.

Year of publication: 2007

Towards a Measure for Arabic Corpora Quality

Yassine Benajiba, Paolo Rosso

نحو مقياس لجودة الذخيرة اللغوية العربية

في الآونة الأخيرة، ساعدت الأبحاث الإحصائية لمعالجة اللغات الطبيعية (NLP) بشكل كبير في تسهيل استخدام الشابكة ومهامها التلقائية التي كانت بديلا للمهام اليدوية التي توصف بالمملة والمكلفة. وعلى وجه التحديد سمحت محركات البحث مثل جوجل (Google) وياهو (Yahoo)، لجميع أنواع مستخدمي الشابكة الوصول إلى الوثائق ذات الصلة وصفحات الويب لاستعلاماتهم. و تسمح المترجمات التلقائية مثل مترجم جوجل أو بنج (Bing) للمستخدمين بالاستفادة من المعلومات والوثائق المكتوبة في اللغات الأجنبية.

قامت هذه الورقة البحثية بإيجاد مقياس إحصائي يستخدم لأول مرة لتقييم جودة الذخيرة اللغوية في اللغة العربية. ويستند هذا المقياس على البيانات الإحصائية ولا يعتمد على اللغة. ومع ذلك، فإن القيم التي يمكن الحصول عليها من التجارب يمكن أن تكون مختلفة جدا بالاعتماد على الذخائر اللغوية المكتوبة بلغات مختلفة، وقد أجرى الباحثان تجاربهما على اللغة العربية تحديدا، فاختارا أربعة أنواع مختلفة من الذخائر اللغوية من أجل تحديد خصائص كل نوع منها، ويعكسها المقياس الذي اعتمده الباحثان لقياس الجودة. أظهرت النتائج الأولية أن المقياس يرتبط ارتباطا وثيقا بأسلوب الكتابة وطبيعة النص الذي تم اختياره.

احتوى النوع الأول على أكثر من ٦٦,٠٠٠ كلمة (هذا يعادل ٣٦٠ كيلو بايت) من شعر أبي الطيب المتنبي. أما الذخيرة اللغوية الثانية فقد تم الحصول عليها من مجموعة مقالات إخبارية مكونة من أكثر من ٥٠,٠٠٠ كلمة (هذا يعادل ٢٦٠ كيلو بايت) من واضيع مختلفة.

أما بالنسبة للذخيرة اللغوية الثالثة، فقد أخذت من كتاب علمي تعليمي وعنوانه (A Linux Red Hat) إذ اختير من أجل دراسة التدابير التي يمكن الحصول عليها للذخيرة اللغوية، وقد تكونت هذه المجموعة من أكثر من ٥٥,٠٠٠ كلمة (ما يقارب ١٢٦ كيلو بايت). احتوت الذخيرة اللغوية الأخيرة على مقتطفات من كتاب ديني للإمام ابن القيم الجوزية، احتوت هذه المقتطفات على ما يقارب ٦٥,٠٠٠ كلمة (أكثر من ٤٦٠ كيلو بايت). قبل القيام بعملية حساب عوامل القياس لكل من الذخائر اللغوية الأربعة، قام الباحثان بتجزئة النص كخطوة أولية، أي المعالجة المسبقة. لهذا الغرض قاما باستخدام برنامج التحليل المتاح مجانا على المواقع.

قدّم هذا البحث تعريفا لمقياس الجودة للذخائر العربية الذي تدعمه بعض التجارب الأولية. ويتكون من ثلاثة عوامل رئيسة، وهي كالآتي: التعقيد الذي يعكس عدد الكلمات والجمل المستخدمة في الذخيرة اللغوية، وقد لوحظ أن هذا العامل يزداد أهمية في البيانات التي تركز على المحتوى أكثر من نمط الكتابة. العامل الثاني يعتمد على التنوع الذي يبين التفاوت والاختلاف في المفردات المستخدمة في البيانات، وكانت نسبتها مرتفعة أيضا في الأصناف العلمية. أما العامل الثالث فيعتمد على توزيع تكرار الكلمات (words frequency dis tribution) المرتبط بنوعية الكتابة المستخدمة في البيانات وأسلوبها.

Learner Corpus Studies in Asia and the World (LCSAW) 2014, May 31 - June 1, 2014, Kobe University, Japan. Kobe International Communication Center, 77 - 89.

Year of publication: 2014

– Arabic Learner Corpus (ALC) v2

A New Written and Spoken Corpus of Arabic Learners

Abdullah ALFAIFI, Eric ATWELL, Hedaya IBRAHEEM

الذخيرة اللغوية لدارسي اللغة العربية "ALC (v2)" – ذخيرة جديدة مكتوبة ومنطوقة لدارسي اللغة العربية

تستخدم الذخيرة اللغوية في التعلم بشكل متزايد في بعض مجالات البحث اللغوي مثل تعليم اللغة وتعلمها، واللغويات التطبيقية (applied linguistics)، والصناعة المعجمية (lexicography)، وكذلك تستخدم لأغراض أخرى مثل تحليل الأخطاء (error analysis)، ومراقبة تحسن مستوى الدارسين، وتصميم مواد اللغة، وتحليل التباين بين اللغات، وبناء قواميس الدارسين وقواميس الأخطاء الشائعة. ومع ذلك، فهناك افتقار شديد إلى ذخيرة لغوية متاحة بالمجان لدارسي للغة العربية ، وهذا قد يفسر النقص في البحوث التي تركز على اللغة العربية في مجالات البحث اللغوي المذكورة أعلاه. قامت هذه الورقة البحثية باستكشاف مشروع تجميع ذخيرة لغوية لدارسي اللغة العربية التي يجري تطويرها في جامعة ليدز، والمتاحة للاستخدام العام من الباحثين في اللغة العربية، والمتكونة من ((٢٨٢٬٧٣٢ كلمة جمعت من دارسي اللغة العربية في المملكة العربية السعودية. وتتضمن المجموعة بيانات مكتوبة ومنطوقة من إنتاج (٩٤٢) طالباً من (٦٧) جنسية مختلفة يدرسون في مرحلة ما قبل الجامعة ومرحلة الجامعة. كما تهدف الذخيرة اللغوية لدارسي اللغة العربية إلى توفير مورد مفتوح للبيانات لبعض مجالات البحث اللغوي المتعلقة بتعلم اللغة العربية وتعليمها، وبالتالي، فإن بيانات المجموعة

بكاملها متاحة للتحميل من شبكة الإنترنت، و يحتوي الإصدار الحالي على (١٥٨٥) ملفاً.

لم تتلق الذخيرة اللغوية للدارسين اهتماماً كافياً، وخاصة لتعلم اللغة العربية كلغة ثانية في البلدان الناطقة باللغة العربية. واستناداً إلى بعض الأدبيات، فإن هناك عددا قليلا من المشاريع التي تقوم بتطوير ذخيرة لغوية لدارسي اللغة العربية، ولا تتوافر معظمها مجاناً للاستخدامات أو للباحثين، بالإضافة إلى أن تلك المشاريع تهدف إلى المساعدة في اكتساب اللغة العربية كلغة أجنبية؛ حيث جُمعت من دارسي اللغة العربية في البلدان غير الناطقة بالعربية.

استعرضت هذه الورقة البحثية عدداً من الذخائر اللغوية التعليمية ذات الصلة، كما أظهرت معايير تصميم الذخيرة اللغوية لدارسي اللغة العربية التي شملت كلاً من اللغة المستهدفة، والمشاركين، وحجم المجموعة، والمواد المدرجة، والطريقة والمهام المستخدمة لجمع البيانات فضلاً عن البيانات الوصفية لكل من المواد الأساسية للذخيرة والمساهمين فيها. كما قُدمت تفاصيل حول محتوى الإصدار الحالي من الذخيرة اللغوية لدارسي اللغة العربية استناداً إلى (٢٦) عنصراً تمثل البيانات الوصفية للذخيرة التي تتضمن البيانات الوصفية لكل من الدارسين والنصوص، كما جرى القيام بمزيد من العمل للتعليق على الذخيرة، فعلى سبيل المثال، عُملت نسخة ثانية من علامات الخطأ للذخيرة اللغوية لدارسي اللغة العربية مع دليل إضافة علامات الخطأ. بالإضافة إلى ذلك، يجري حالياً وضع أداة لإضافة هذه العلامات بطريقة محوسبة.

7th international conference on language engineering, Cairo (2007)

Year of publication: 2007

Building an International Corpus of Arabic (ICA): Progress of Compilation Stage

Sameh Alansary, Magdy Nagi, Noha Adly

بناء الذخيرة العالمية للغة العربية "ACI" : تقدم مرحلة التجميع

دعت الحاجة إلى بناء الذخائر اللغوية لسببين رئيسين، الأول أنه لا يمكن للذخائر اللغوية مهما كانت كبيرة أن تحتوي على معلومات حول جميع المجالات اللغوية وذلك لاختلاف المعاجم لغويا أو اختلاف مجالات قواعد تلك اللغة. والسبب الثاني أن كل ذخيرة لغوية مهما كانت صغيرة قد تقوم بتعليم الإنسان حقائق لا يمكن تصور تعلمها بأية طريقة أخرى. وهكذا فقد جُمع عدد كبير من الذخائر اللغوية في السنوات القليلة الماضية، لكن الفكرة نفسها ليست جديدة. ويمكن أن ترجع الفكرة إلى اللغوي الألماني (كادنج) الذي قام في عام ١٨٩٧، باستخدام ذخيرة لغوية كبيرة للغة الألمانية تضمنت أحد عشر مليون كلمة لجمع توزيعات الترددات من الحروف وتسلسل الحروف.

تتوافق أهمية الذخائر اللغوية في الدراسات اللغوية مع أهمية البيانات التجريبية، فمن المهم جداً معرفة أن البيانات التجريبية في اللغة والبحوث اللغوية ضرورية، نظراً لأن هذا النوع من البحوث لا يمكن أن يعتمد على الإدراك المعرفي الفردي الداخلي (**individual's own internalized cognitive perception**). وفق هذا الرأي، فإن التحقيقات التي تستند إلى الذخيرة مفيدة جداً؛ وهذا يمكن أن يكون واضحاً جداً في الدراسات المعجمية وقواعد اللغة ومعالجة اللغات الطبيعية (**NLP**) والعديد من الدراسات اللغوية الأخرى.

تركز هذه الدراسة على ثلاثة محاور: المحور الأول يعطي استطلاعاً لأهمية الذخائر اللغوية في دراسات اللغة، مثل قواعد اللغة ومعالجة اللغة الطبيعية والصناعة المعجمية

(Lexicography) وعلم دلالات الألفاظ (Semantics) وغيرها من المجالات. أما المجال الثاني فيوضح افتقار اللغة العربية إلى موارد نصية مثل الذخائر اللغوية وأدوات تحليلها، وتأثير هذا النقص على جودة تطبيقات اللغة العربيةٍ وندرة وجود تجارب ناجحة في مجال تجميع الذخائر اللغوية العربية، وبالتالي فإن المحور الثالث يعرض التصميم التقني للذخيرة الدولية للغة العربية (International Corpus of Arabic (ICA))، وهي عبارة عن ذخيرة تم إنشاؤها حديثاً كتمثيل لغوي للغة العربية، وكان الهدف من ورائها تغطية اللغة العربية ليجري استخدامها في جميع أنحاء العالم العربيٍ ومن المقرر أن تدعم هذه الذخيرة مجموعة من الدراسات العربية التي تعتمد على البيانات الأصلية، بالإضافة إلى بناء تطبيقات معالجة اللغة العربية.

كما تُقدم هذه الورقة الوضع الحالي للذخيرة الدولية للغة العربية(ICA) من حيث تصميمها والبرنامج الأولي المستخدم في استجواب الذخيرة. ويمكن اعتبار هذه التجربة واحدة من النماذج الناجحة لبناء ذخيرة تمثيلية للغة العربية الحديثة (MSA). ومن المهم إدراك أن إنشاء ذخيرة دولية للغة العربية (ICA) يجب أن يكون عملية مستمرة ، تتطلب إعادة تقييم مستمرة خلال فترة تجميع الذخيرة. ومن المهم أيضاً التأكد من أنواع الموارد ، والموارد الفرعية، والأنواع، والأنواع الفرعية، التي يتعين إدراجها في الذخيرة الدولية للغة العربية (ICA). وبمجرد الانتهاء من عملية جمع النصوص والحوسبة، ستكون النصوص جاهزة للمرحلة النهائية من الإعداد، وبعد ذلك سيكون من السهل التعامل مع النصوص في مرحلة التحليل.

International Journal of Computational Linguistics Research Volume 6 Number 3 September 2015

Year of publication: 2015

Corpus-Based Prediction of Coordination Ambiguity in Arabic

Wafaa Daffa, Raad Alshahry, Imtiaz Hussain Khan

التنبؤ بغموض التنسيق في اللغة العربية المستند إلى ذخيرة لغوية

الغموض هو سمة من سمات النص، حيث يمكن تفسير النص عادة بأكثر من طريقة مختلفة، ففي اللغات الطبيعية مثل العربية والإنجليزية يمكن أن ينشأ الغموض على مستويات مختلفة. أكثر أنواع الغموض شيوعاً هو الغموض الهيكلي (Structural Ambiguity) الذي يعرف أيضاً باسم الغموض النحوي (Syntactic Ambiguity)، إذ يمكن ترتيب سلسلة من الكلمات بشكل نحوي بأكثر من طريقة مما يؤدي إلى أكثر من تفسير واحد. وهناك شكل نموذجي من الغموض الهيكلي هو غموض التنسيق (Coordination Ambiguity)، يعمل كمعدّل خارجي (External Modifier) في هياكل منسقة مثل عبارة «القطط والكلاب السوداء» التي تعد عبارة غامضة، لأن القارئ قد يميل إلى تفسير هذه العبارة على أن كلا من القطط والكلاب سوداء، أو أن الكلاب فقط سوداء. ويفترض، في هذه الحالة، أن أغلبية القراء قد تختار التفسير الأول. ومن المهم أن يذكر هنا أن الصفات في اللغة العربية تتبع الموصوف في الصفات كالتذكير والتأنيث مثلا.

تشير الأدلة النفسية اللغوية إلى أن هذا الغموض يسبب في كثير من الحالات التباساً في المعنى. وعلى الرغم من أن العبارة من الناحية النظرية قد تكون غامضة، إلا أن معظم الناس قد يفسرون العبارة بنفس الطريقة. لذلك فإن المشكلة هي كيف يمكن لنظام حاسوبي تحديد احتمال تفسيرات مختلفة لعبارة معينة ومن ثم تحديد التفسير الأرجح. بالمقارنة مع الأشكال

الأخرى من الغموض الهيكلي، لم يحظ غموض التنسيق إلا بالقليل من الاهتمام في الأدب العربي. في هذه الورقة البحثية، تناول الباحثون مشكلة إزالة غموض هياكل التنسيق في اللغة العربية لتحديد كيفية تطبيق المعدِّل الخارجي (external modifier) على الكلمات أو العبارات المنسقة. يمكن القول بأن الكلمات والعبارات من جميع الأنواع يمكن تنسيقها. ومع ذلك، لدراسة بيانات محددة، فقد ركز الباحثون على عبارات الأسماء الغامضة المحتملة من النوع الذي يحتوي على اسم أول واسم ثان وصفة. لقد قدر الباحثون التفسير الأرجح الذي قد يتبناه القراء البشر لعبارات الأسماء الغامضة المحتملة باستخدام معلومات إحصائية عن حدوث المشترك المعجمي (Lexical co-occurrence).

وقد بحثت الدراسة التجريبية في هذه الورقة إمكانية استخدام بيانات تجميع الكلمات المستندة إلى قاعدة بيانات للتنبؤ بالتفسيرات المختلفة المحتملة لجملة غامضة في اللغة العربية. وكدراسة حالة، فقد عالج الباحثون مشكلة إزالة غموض هياكل التنسيق في اللغة العربية لتحديد كيفية تطبيق المعدِّل الخارجي أو الصفة على الكلمات المنسقة أو الأسماء. وقد بنيت التجارب بطريقة تتسم بالتوزيع المتوازن للصفات ذات التجميع الإحصائي العالي أو المنخفض للكلمات مع الاسمين. وقدمت إلى عدد من المشاركين المتكلمين بالعربية سلسلة من التجارب التي تحتوي كل تجربة منها على جملة غامضة محتملة يليها سؤال للفهم له علاقة بالجملة السابقة. وكشفت البيانات أنه يمكن استخدام المعلومات المعجمية المشتركة (lexical co-occurrence information) للتنبؤ بالتفسير الأكثر ترجيحاً لجملة غامضة محتملة.

8th International Conference on Language Engineering, Egypt, Pages

Year of publication: 2008

Towards Analyzing the International Corpus of Arabic (ICA) : Progress of Morphological Stage

Sameh Alansary, Magdy Nagi, Noha Adly

نحو تحليل الذخيرة الدّولية العربية "ICA": تطوُر مَرحلة الصّرف

المُقدمة

يُسلّط هذا البحث الضوء على أربعة محاور: المحور الأول، يتعامل مع مستويات تحليل الذخيرة، مثل التحليل الصرفي؛ والتّحليل المعجمي (معاني المفردات)؛ والتحليل النّحوي؛ والتّحليل الدلالي. أما المحور الثاني، فيُناقش بعض محاولات تحليل الذخائر اللغوية العربية. في حين يُوضح المحور الثالث مختلف الأدوات المتاحة للتحليل الصّرفي العربي (وهي Xerox, Tim Buckwalter, Sakhr، RDI) وأخيراً يتعامل المحور الرابع، وهو القسم الأساسي في هذه الورقة، مع التحليل الصّرفي للذخيرة اللغوية العربية ICA، التي تشتمل على: اختيار نموذج التّحليل ووصفه (analysis model)، ومرحلة التحليل الأولي (المسبّق) (pre analysis model) ومراحل تحليل النّص الكامل.

المبررات ومشكلة البحث

يعتمد التحليل اللُغوي للذخيرة إلى حدٍ كبير على توافر تجارب تاريخية سابقة في التحليل، حيث تُمثل المعلومات مرحلة أولى لتوفير الحُلول الحاسمة، وتُستخدم أيضاً في المراحل التالية من التحليل. والفرق الرئيسي بين إنشاء الذخيرة وتحليلها هو أنه في حين أن مُنشئ الذخيرة لديه خيار تعديل ما هو مُدرج في الذخيرة للتخلُص من أي تعقيدات قد تظهر خلال إنشاء الذخيرة فإن محلل الذخيرة يواجه متناً ثابتاً جامداً، وعليه أن يقرر ما إذا كان ينبغي الاستمرار في تحليله،

حتى لو كانت الذخيرة غير مناسبة تماماً للتحليل، أو أن يُحاول العثور على ذخيرة جديدة.

منهجية البحث

يتضمّن التّحليل اللُغوي أكثر من مُستوى من مُستويات التّحليل، مثل التّحليل الصّرفي، والتّحليل المعجمي، والتّحليل النّحوي والتّحليل الدلالي. وعادةً ما يكون التركيز في تحليل الذخيرة على الجانب التجريبي، في حين أن التفسير يمكن أن يكون نَوعياً أو كميّاً. ويعدّ التحليل الصرفي أبسط أنواع التحليل اللغوي، لأنه يشكل الأساس لأنواع أخرى من التحليل (مثل التحليل النحوي والشُروح الدلالية). إنّ الهدف من التحليل الصّرفي للمتن ليس فقط تعيين رمز أو علامة لكل وحدة معجمية في النص تشير إلى جزء من الكلام، لكن أيضاً الإشارة إلى معلومات أخرى صرفية.

الخاتمة

تُمثّل هذه الورقة خارطة طريق لمحاولة تحليل الذخيرة العربية، حيث اتبع المحللون أسلوباً قائماً على التّجذيع (أو استخراج جذع الكلمة) لاستخدامه في تحليل الذخيرة الدولية العربية ICA . ويُعدّ مُحلل (Buckwalter) – من بين المحللات المعجمية المتاحة – أفضل المحللات الصّرفية وأنسبها على الإطلاق لأسلوبنا.

لقد ناقشت هذه الورقة عددًا من الاعتبارات العامة التي يجب مراعاتها عند بدء عملية تحليل الذخيرة الدولية العربية ICA. ويمكن اعتبار هذه التجربة واحدة من أنجح منهجيات تحليل اللغة العربية الحديثة بالمقارنة مع التجارب الأخرى لتحليل الذخائر العربية. وسيتم تطوير العيّنة التي تم تحليلها لغرض استخدامها كذخيرة تدريبية لتحليل الحجم المستهدف من الذخيرة العربية ICA والبالغ (١٠٠ مليون كلمة). وسيتم تطوير برنامج ICA لاختبار النُسخة التي حُللت بشكل كامل لمساعدة الباحثين في الحُصول على بحث نصّي قوي.

Proceedings of the Second International Conference on Arabic Language Resources and Tools

Year of publication: 2009

PROCESSING LARGE ARABIC TEXT CORPORA: PRELIMINARY ANALYSIS AND RESULTS

Fahad A. Alotaiby, Ibrahim A. Alkharashi, Salah G. Foda

مُعالجة ذخائر النُصوص العربية الكبيرة: تحليل أولي مع النّتائج

المقدمة ومُشكلة البحث

تعتمد العديد من مجالات البحث المهمة مثل التعرُف التلقائي على الكلام أو الخِطاب المنطوق، والتعرف الضّوئي أو البصري على الحُروف، واسترجاع المعلومات اعتماداً كبيراً على وجود نموذج للغة (Language model) أو نموذج إحصائي (Statistical Model) جيد عن اللُغة المستخدمة. ويؤدي التمثيل الأكثر دقة إلى نشوء أو تطوير أنظمة أكثر دقّة. من جهةٍ أخرى، تُعدّ اللغة العربية أكثر ثراءً وأكثر تعقيداً بكثير من اللُغة الإنجليزية، وهذا ما أبرز الحاجة إلى دراسة الإحصاءات الرئيسية للُغة العربية، والفروق الإحصائية بين اللغتين العربية والإنجليزية على نطاق واسع.

هي الخطوة الأولى لمعالجة أي نص أو ذخيرة ولعلّ تقسيم النص المدخل إلى وحدات. ويمكن أن تكون هذه الوحدات عبارة عن أحرف أو كلمات أو أرقام أو جُمل أو أي وحدة أخرى مناسبة. ولا يكون تعريف الكلمة هنا مطابقاً على نحوٍ دقيق للشكل أو التركيب النحوي، وهذا هو السبب في أننا نسميها «رموزاً» أو «علامات».

منهجية الدراسة

ولأغراض هذه الدراسة، جرى استخدام اثنين من الذخائر العربية والإنجليزية الكبيرة

والشاملة، وهما عبارة عن «الطبعة العربية الثالثة من جيجاوورد» و«الطبعة الإنجليزية الثالثة من جيجاوورد» على التوالي. في هذه الورقة، قمنا باستخدام هاتين الذخيرتين لإجراء تحليلنا الأولي وعرض نتائج اللغة العربية بالاقتران مع اللُغة الإنجليزية. هدفت هذه الورقة إلى تقديم إحصاءات حول توزيع الرموز أو الدلالات وطول الفقرات، وعلامات الترقيم، ومُفردات اللُغتين العربية والإنجليزية. كما ناقشت هذه الورقة اعتبارات وقضايا المعالجة الأولية.

النتائج والتوصيات

في هذا البحث استعرضت الاختلافات الإحصائية بين اللُغتين العربية والإنجليزية على نطاق واسع. وتم عرض نتائج استخدام ذخيرتين باللغتين العربية والإنجليزية مأخوذتين من ذخيرة «جيجاوورد»، وتم الحصول على ٦٠٠ مليون كلمة باللغتين العربية والإنجليزية. وأظهرت نتائج الدراسة أن عدد أنواع الكلمات العربية أكثر بنسبة ٧٦٪ من اللُغة الإنجليزية. ويمكن تفسير ذلك لأن اللغة العربية أكثر ثراء من اللُغة الإنجليزية. كما استُعرِضت نتائج التوزيعات الإحصائية لطول الكلمة وطول الفقرة في كلٍ من اللغتين العربية والإنجليزية. من جهةٍ أخرى، أظهرت الدراسة بأن الوثائق العربية تعاني من أخطاء هجائية، وسوء استخدام لعلامات الترقيم وإهمال التّنسيق. لذلك، قد تكون المعالجة المسبقة لمثل هذه الوثائق خطوة مهمة قبل استخدامها.

٣-٣-٢ أبحاث المدونات الموسومة

وتضم خمسة أبحاث، بينها ثلاثة أبحاث نوع أ هي: **منهجية الذّكاء الاصطناعي في معالجة المحتوى العربي والإسلامي على الشّابكة، و مدوّنة «بِن» العربيّة المشجرة - بناء مدونة عربيّة موسومة ذات نطاق واسع، و قائمة علامات الوَسم الصّرف-نَحوي لنصوص اللغة العَربيّة.**

أما الأبحاث نوع ب فكان هناك بحثان هما: **بناء نظام وسم تلقائي متكامل للنصوص العربيّة، و الوسم النحوي يدويا للغة القرآن الكريم.**

NITS 2011 3rd National Information Technology Symposium, 6-9 March 2011, Riyadh, Saudi Arabia.

Year of publication: 2011

An Artificial Intelligence approach to Arabic and Islamic content on the internet

Eric Atwell, Claire Brierley, Kais Dukes, Majdi Sawalha, Abdul-Baquee Sharaf

منهجية الذّكاء الاصطناعي في معالجة المحتوى العربي والإسلامي على الشّابكة

الملخص:

تستعرض هذه الدراسة مجموعة من الأبحاث التي تناقش منهجية الذكاء الاصطناعي وعلم لُغويات مُتون اللغة العربية والقرآن الكريم، في جامعة ليدز. وخلُصت هذه الأبحاث إلى إنتاج العديد من البَرمجيّات وقوائم بيانات المتون لأغراض بحوث «اللغة العربية الحديثة» بُحوث لغة القرآن. لقد نال عملُنا في اللغويات القرآنية اهتماماً واسعاً، ليس فقط من جانب اللغويين العرب لكن أيضاً من الطّلبة الدّارسين للقرآن؛ وحتّى من عامة الناس. ونرى اليوم الأثر الكبير الممكن لنَمْذَجة الذّكاء الاصطناعي للقرآن الكريم، وهو ما يقودُنا إلى اقتراح مزيد من الأبحاث حول خارطة المعرفة القُرآنية.

المقدمة

شاركت مجموعة البُحوث الخاصة بمعالجة اللغات الطبيعية، وهي مجموعة تابعة لمعهد الذكاء الاصطناعي والنظم البيولوجية **I-AIBS** في جامعة ليدز، في بحثٍ خاص بمُعالجة اللغة العربيّة الطبيعية وعلم لغويات المتون لفترة تزيد على عقد من الزّمان. وكان تركيزُنا في المراحل الأولى من البحث على الأدوات و الموارد الخاصة بعلم اللغة لتحليل «اللغة العربية الحديثة» ونَمذَجتها. ثمّ تناولنا في وقت لاحق من البحث «اللغة القرآنية». ونحن ننظر للقرآن الكريم باعتباره مجموعة غَنيّة من البيانات التي تغني الذكاء الاصطناعي وبُحوث التّعليم الآلي.

ويُعد النص القرآني في العقيدة الإسلامية نصاً فريداً من نوعه؛ فهو كلام موحى به من عند الله، بواسطة الوحي جبريل الى النبي محمد، وجاء تدوينُه حرفياً ليكونَ المصدر الموثوق والوحيد للمَعرفة والحِكمة والتّشريع.

مشكلة الدراسة

يُواجه الباحثون في مجال الذّكاء الاصطناعي بعض التّحديات حول كيفية تقديم تلك المعرفة والحكمة والتشريع من خلال أنظمة حاسوبية لبِناء أنظمة ذكية بإمكانها الإجابة عن أية تساؤُلات تخُصّ المعرفة من النّص القرآني، ومساعدة المجتمع الإسلامي وغير الإسلامي على فهم القرآن وتدبّره.

وبشكل عام، تَعتمد بُحوث الذكاء الاصطناعي على مَتْن اللغة (وهو عبارة عن قائمة بيانات النّص – مَدار البحث – والغني بِمجموعة من البيانات الوصفية والعلامات أو الشُروح التي تُبيّن التّحليلات الصّرفية وعلامات أقسام الكلام وغيرها. وهناك الكثير من الأمثلة على ذلك لا يتّسع المقام لحَصرها.

ويضم القرآن الكريم قائمة من البيانات النّصيّة التي لا تزيد الكلمات المكوّنة لها عن ٨٠,٠٠٠ كَلِمة مُتسلسِلة ومُرتّبة ضِمن الأجزاء والآيات القُرآنية. ويعتقد المسلمون بأنّ الصّيغة الأصلية المنطوقة لهذه البيانات هي اللغة العَربية التّقليديّة (الكلاسيكيّة)، التي جرى تدوينُها بأمانة في نظام نسخ مُتطور. ولقد دأب العُلماء عبر ما يزيد على الألف سنة على البحث حول كيفية استخلاص المعرفة والتّشريعات من هذا النّص القرآني، كما قاموا بتصنيف العَديد من التفاسير أو المتون التّحليليّة والشُروحات.

وقد ناقشت هذه الدراسة المحاور الرّئيسية التّالية:

أولاً: بُحوث الذّكاء الاصطناعي على اللغة العربيّة والقُرآن الكريم في جامعة ليدز

إنّ أنظمة استرجاع المعلومات واستخلاص البيانات موجودة فعليّاً، مما يتيح لعُلماء القرآن

الحُصول على النُصوص القُرآنية من خلال عدد من المواقع على الشّابكة؛ إذاً لماذا يُعدّ فَهم القُرآن تَحدياً كبيراً للذّكاء الاصطناعي؟

ثانياً: نحو مزيد من البُحوث حول: خارطة المَعرفة القُرآنية

من أجل توسيع نطاق البحث عن المحتوى العربي والإسلامي على الشّابكة، تقدّمنا بمُقترح مشروع حوسبة يمثل تحدياً وهو (خارطة المعرفة القرآنية)، وهو مورد غَني مُنظّم ودقيق، يكون متاحاً على الشّابكة، لفهم القرآن الكريم والنصوص الدينيّة للإسلام. وسوف يكون هذا النّظام بمثابة قاعدة بيانات مُنظّمة وقابلة للقراءة من المعلومات اللغويّة والدلالية لتمهيد الطّريق أمام المزيد من البحث، فضلاً عن كونِه مَوقعاً تَعليمياً مُفيداً للغاية. وهناك مُبررات قويّة جداً وحاجة متزايدة لوجود مورد على الشّابكة لمعرفة قرآنية عالية الجَوْدة.

الحُصول على المَعرفة القُرآنية

ذكرنا آنفاً حجم التّشويه الذي تعرّضت له المعرفة القُرآنيّة والمفاهيم الإسلامية في وسائل الإعلام الغربي وعبر الشّابكة عموماً. وهو ما أظهر الحاجة لطرح مورد خبرة موضوعي ومُنصف – عبر الشّابكة – لإعطاء المجتمعات حول العالم صورة مُختلفة حول معنى الاسلام الحقيقي. ومن هنا، يعدُّ فهم القُرآن الكريم التّحدي الأعظم أمام علوم الكمبيوتر والذكاء الاصطناعي.

كما أنّ بعض المسلمين من غير النّاطقين باللغة العربية قد يجهلون المعاني العميقة للقرآن رغم تذكرهم أصوات الآيات القرآنية، وأمثال هؤلاء قد يساعدهم وجود خارطة المعرفة القُرآنية على تعلّم المزيد والحصول على إجابة عن تساؤلاتهم حول القرآن.

تطوير متن عربي قُرآني

إنّ المتن العربي القرآني هو سلسلة من قوائم البيانات اللغوية المخططة التي تمّ تطويرها في جامعة ليدز، من خلال الشرح الجماعي عبر الشابكة في جميع أنحاء العالم. وحتى الآن، تم

إطلاق مجموعات البيانات الصّرفية والنّحوية، والعمل جارٍ على تطوير توصيف اللغة الدلالية للقرآن الكريم.

تصميم الوحدات القياسية

من المنهجيات الرئيسية في تطوير خارطة المعرفة القرآنية: بناء سلسلة منظمة من الوحدات القياسية ذات الصّلة، التي تتضافر معاً لتُكوّن المشروع النهائي. وقد لخّصت الدراسة الوحدات القياسية التي يتكوّن منها النموذج المقترح للخارطة كالآتي:

* البنية التحتية
* قوائم البيانات
* تطبيقات المُستخدم النهائي

وأخيراً، من المتوقع أن يستفيد المشروع من مجموعة كبيرة من الخبراء المتطوعين حول العالم من خلال الشروح التعاونية (المشتركة)، وهي منهجية مُجربة تُستخدم لتطوير اللغة العربية القرآنية، وذلك إذا تم وضع الإطار العملي الصحيح والبنية التحتية المناسبة.

NEMLAR Conference on Arabic Language Resources and Tools, pages 102–109, Cairo, Egypt

Year of publication: 2004

The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Wigdan Mekki

مدوّنة "بِن" العربيّة المشجرة - بناء مدونة عربيّة موسومة ذات نطاق واسع

تستعرض هذه الورقة – اعتمادا على ثلاث سنوات من الخبرة في تطوير مدونة ذات نطاق واسع من النصوص العربية الموسومة (**Annotated Arabic text**) – القضايا الآتية: (أ) اختيار المنهجية المناسبة لمعالجة إشكاليات حوسبة اللغة العربيّة (ب) شرح أسباب اختيارنا أسلوبَ مدونة بن المشجرة (**Penn Treebank**) الإنجليزي في وسم المدونة (التي تتطلب من المدونين الناطقين بالعربيّة التعامل مع نظام نحوي جديد لوسم المدونة) بدلا من استخدام نظام النحو التقليدي لقواعد اللغة العربيّة (ج) عرض العديد من النماذج التي تظهر أهمية الوسم اليدوي المبني على الفهم البشري في مقابل صعوبة التحليل الآلي وخاصة عند التعامل مع غموض المكتوب (**Orthographic ambiguity**) (د) إعطاء مثال توضيحي على منهجية مدونة بن العربيّة المشجرة، ومتابعة عمليات الوسم والتحليل الصرفي والوسم والتحليل النحوي بالتفصيل، (هـ) الختام بما تحقق حتى الآن وما يجب القيام به.

المقدمة

تحقق على مدى العقد الماضي تقدم مهم في المعالجة الحاسوبية للغة العربيّة. بيد أنه نظرا لواقعها الاجتماعي–السياسي، وتعقيد بنائها الصرفي، والاختلافات الكبيرة بين لهجاتها، ما تزال اللغة العربيّة تشكّل تحديا للعاملين في حوسبة اللغات. تفتقر إلى أدوات الحوسبة والنصوص الموسومة، حتى أنها لا تمتلك برنامج تقطيع العبارة (**Base Phrase Chunkers**) وفي المقابل، هناك زيادة في الطلب على موارد حوسبة اللغة العربيّة عالية الجودة.

يتفق العاملون في معالجة اللغات الطبيعية وتكنولوجيا اللغات البشرية (Human Language Technology) في الأكاديميا وفي الصناعة على أن المدونات المشجرة، ومكملاتها من مدونات المفاعيل (Proposition Banks)، والمعاجم ثنائية اللغة (Bilingual Lexicons)، ومدونات النصوص المتوازية (Parallel Texts) هي أكثر الموارد اللغويّة (Linguistic Resources) استخداما ولزوما في مجالات البحث والتطوير لمعالجة اللغات الطبيعية واستخراج المعلومات وتلخيصها. وتقع المدونات المشجرة ومدونات المفاعيل التي يطلق عليها معا (X-banks) في مركز الأنشطة والتقنيات والمنهجيات التي تقوم عليها أتمتة عملية استخراج المعلومات وفهمها من النصوص.

بدأ العمل في مدونة بن العربيّة المشجرة في خريف ٢٠٠١ وصدر منها حتى الآن ثلاثة إصدارات موسومة صرفيا ونحويا. وقد اعتمدنا في تصميم نظام الوسم للغة العربيّة على القواعد العربيّة النحوية التقليدية، والنظريات النحوية الحديثة، ونهج وسم مدونة بن المشجرة الإنجليزي، الذي نعتقد أنه قابل للتعميم عند تطوير أنظمة وسم للغات أخرى. وقد لجأنا عند الضرورة للابتكار فيما يتعلق بقواعد اللغة العربيّة التقليدية وخاصة عندما كان الوسم النحوي المبتكر معبرا عن المقصود.

ما هي اللغة العربيّة الحديثة (Modern Standard Arabic MSA)

«اللغة العربيّة الحديثة» هي اللغة التي تستهدفها عمليات معالجة اللغة العربيّة حاسوبيا، وقد أوجد العاملون في حوسبة اللغة العربيّة هذا التعبير واستخدموه لمدة طويلة دون الاتفاق على تعريفه. فهي ليست اللغة الأصلية أو الأولى لمجتمع، إلا أنها لغة الخطاب المكتوب، وتستخدم في الاتصالات الرسمية سواء كانت مكتوبة أو شفهية، ولها مجموعة محددة من الأساليب. ويمكن لمصطلح «العربيّة المكتوبة الحديثة»، (Modern Written Arabic) أن يكون أكثر ملاءمة من «اللغة العربيّة الحديثة» لولا الغموض والجدل الذي سيلحق به لدخول اللهجات المحكية فيما يكتب اليوم. المصطلح الآخر الذي ظهر في أدبيات معالجة اللغة العربيّة حاسوبيا هو «لغة

الحوار العربيّة الحديثة"، (Modern Conversational Arabic) وعلى الرغم من أنه قد يكون مفيدا ومقبولا في الاصطلاح، إلا أن المشكلة فيه أنه يفتقر إلى خصوصية ربطه بلهجة واحدة محددة من طيف اللهجات العربيّة الممتدة والمتداخلة لغياب معيار يساعد في هذا التحديد.

أثر خصوصية اللغة العربيّة على وسم المدونة

يمكن تلخيص أثر خصوصية اللغة العربيّة على عملية وسم المدونة ومنهجية الوسم في النقاط الآتية:

١- غياب التشكيل عن معظم النصوص العربيّة المكتوبة، واقتصاره على عدد قليل من النصوص الدينية والمدرسية.

٢- يبدو أن معظم تطبيقات معالجة اللغة العربيّة لا تتعامل مع تشكيل الكلمة، وهو ما يفقدها العديد من السمات اللغويّة الهامة وخاصة النحوية منها.

٣- يقرأ القارئ النص العربي ويفسر معناه عقليا لتعويض المعلومات النحوية المفقودة بسبب غياب التشكيل، وهو ما يؤدي إلى درجة مقبولة من فهم النص. وهـذه العملية ذات أثر غير قليل في اتخاذ القرارات الدقيقة والسريعة عند وسـم النص يدويا.

٤- إن وجود علامات التشكيل ليس ضروريا تماما لوسم المدونة لأن العمل يدوي وليس آليا. ومع ذلك، فإن وجود هذه العلامات يكمل النص ويعزز نوعية التحليل اللغوي. وفي القراءة الأولى يقوم المتخصص بالصرف بتفسير النص الخالي من الحركات حسب معرفته اللغويّة، وينتج نصا مشكولا جاهزا للتحليل النحوي بناء على فهمه. وعند القراءة الثانية التي يقوم بها المتخصص بالنحو فله أن يقبل تفسير المتخصص بالصرف أو أن يرى تفسيره الخاص ويسم النص بناء عليه.

٥- إن وجود نصوص مشكولة يزيل الغموض الناجم عن نقص العلامات النحوية والمعجمية. ومع ذلك، ينبغي أن يكون واضحا أن اللغة العربيّة، شأنها شأن أيّ لغة أخرى، قد تحتوي نصوصها قدرا من الغموض اللغوي.

٦- وبما أن القارئ يفسّر النص بناء على مدى معرفته بالصرف والنحو، فقد اختلف الوسم في بعض الأحيان بين قارئ وآخر. كما اختلف الوسم أحيانا بسبب الاختلافات الناشئة عن التناقض النحوي في النظر للمسألة الواحدة، مما أثر سلبا على الاتفاق بين القراء وعلى جودة عملية الوسم وثباتها.

الخيارات المنهجية

البيانات

اعتمدت المدونة على نصوص الأخبار المتوفرة على شبكة الإنترنت وذلك لوفرتها وحداثتها واستمرار تدفقها، ووجودها بصورة إلكترونية جاهزة لا تحتاج إلى إدخال، بالإضافة إلى عدم وجود قيود تتعلق بحقوق الملكية الفكرية على استخدامها لأغراض البحث العلمي. وقد جمعت بيانات المدونة من مواقع وكالة الأنباء الفرنسية وصحيفتي الحياة والنهار.

اختيار منهجية الوسم للمعلومات الصرفية

اعتمدنا مخرجات عملية التحليل الآلي لمحلل بكوالتر الصرفي للوسم المبدئي للكلمات صرفيا ونحويا. إن هذا المحلل الصرفي يحتوي على قاعدة بيانات للكلمات العربيّة. ويخزن لكل كلمة مكوناتها الصرفية ونوعها النحوي والمعنى المقابل لها في اللغة الإنجليزية بالإضافة إلى ضبطها بالشكل، وكل هذه المعلومات مخزنة بحروف لاتينية. وخلال الفترة من ٢٠٠٢ إلى ٢٠٠٤ قام هذا المحلل الصرفي بوسم ما يزيد على نصف مليون كلمة تحتويها المدونة. ونشرت المدونة الموسومة في ثلاثة إصدارات متتابعة خلال تلك الفترة. وبعد كل إصدار كان يعاد تدوير كلمات المدونة غير الموجودة في المحلل أصلا لتغذيته بها وزيادة كفاءته ومدى تغطيته للكلمات التي يحللها. وقد وصلت درجة تغطيته بعد الإصدار الثالث إلى ٩٩٪ من كلمات المدونة.

وعند إحصاء الكلمات التي سقطت من قاعدة بيانات المحلل الصرفي وجد أن ٣٨٪ منها أسماء أعجمية لأشخاص ولأماكن أو أسماء لشركات ليس لها مقابل في القاموس العربي مثل

أندريوتي أو زيوريخ أو إيرلاينز. وهناك حالات وافقت فيها الأسماء الأعجمية كلمات عربيّة مثل (هو) و(منه) وقد أشكلت على المحلل الصرفي. وفي بعض الحالات لم تحتو قاعدة بيانات المحلل الصرفي في الأصل بعض أسماء الأعلام مثل عادل وأنصاري وبَني وعباد. ثم تأتي أخطاء التشكيل ونسبتها ٢١٪ ومثالها المبني للمجهول. وبعدها الخطأ في تحديد النوع النحوي للكلمة لاحتمالها أكثر من معنى مثل كلمة (عملية) التي تأتي صفة وتأتي اسما. وقد وفر هذا المشروع الفرصة لمراجعة بناء محلل بكوالتر الصرفي ولإغناء قاعدة بياناته مما زاد من كفاءته ومدى تغطيته من الكلمات.

اختيار منهجية وسم المعلومات النحوية

كان التوجه في بداية مشروع مدونة بن العربيّة المشجرة هو اختيار وسوم تستند إلى مسميات تستخدمها قواعد اللغة العربيّة التقليدية بالإضافة إلى استخدام أسلوب وسم مدونة بن الإنجليزية. وذلك لأن فهم النص وفق قواعد اللغة العربيّة التقليدية سيكون أيسر وأكثر ألفة لمن سيضعون الوسوم في النصوص. إلا أن هذا النمط من الوسم سيكون غير مألوف لمعظم الباحثين من غير العرب، وسيكون هناك منحنى تعلم طويل أمامهم ليتمكنوا من استخدام أي بيانات موسومة بهذا النمط التقليدي. أضف إلى ذلك عدم وجود مشاريع سابقة على نطاق واسع تستخدم أسلوب الوسم باعتماد مصطلحات ومسميات قواعد اللغة العربيّة التقليدية، مما يعني تطوير وصقل جميع المبادئ التوجيهية لعملية الوسم من الصفر. ولأن السرعة كانت مهمة لإنجاز هذا المشروع، اخترنا أن نأخذ بمنهجيات الوسم القائمة بالفعل لمدونة بن المشجرة، المعروفة في بيئة البحث العلمي، والمستخدمة في لغات أخرى غير العربيّة. ومع ذلك فقد كان من الضروري إجراء بعض التعديلات عليها لتكون مناسبة للغة العربيّة. وقد أدت بنا هذه التعديلات بالضرورة إلى تعديل عدد من أدوات المعالجة والوسم والتحليل الآلي، إلا أن اعتماد هذا الأسلوب ساعد في الحصول على النتائج بسرعة.

وما زلنا نعتقد أن هذه التجربة كانت فرصة جيدة للمتحدثين باللغة العربيّة ليتعلموا العمل في نظام مدونة بن المشجرة بعد تكييفها لتمثيل بنية اللغة العربيّة.

CL2001: Corpus Linguistics Conference, Lancaster, UK.

Year of publication: 2001

A tagset for the morphosyntactic tagging of Arabic

Shereen Khoja, Roger Garside, Gerry Knowles

قائمة علامات الوَسم الصّرف- نَحوي لنصوص اللغة العَربيّة

المقدمة

مضى نحو أربعة عشر قرناً على ظُهور تقسيم الكلام العربي نحويا، ومع أنّ المبادئ المستخدمة اليوم لتقسيم الكلام مُغايرة قليلاً لتلك المبادئ التي استُخدمت في ذلك العصر، إلا أنه لا يوجد حتى الآن متن عربي موسوم نحوياً. وتتمثل الخُطوة الأولى لشرح متون اللغة في تأليف قائمة علامات (وسوم) يمكنها وصف اللغة وتغطيتها بالكامل وبشكلٍ دقيق. ونظراً لوجود مبادئ لوسم (وضع علامات) اللغة العربيّة، فقد استُخدمت هذه المبادئ لتصنيف علامات الوسم الموصوفة في هذا البحث.

إنّ علامات الوسم العربيّة التي نصفها هنا لا تتبع توصيات **EAGLES** حول وسم المتون الصّرف-نحوية، لكنّ هذا الأمر مُتوقّع لأن اللغة العربيّة تختلف اختلافاً كبيراً عن اللغات التي صمّمها **EAGLES**، فهي لغة ساميّة وليست من اللغات الهندو أوروبية. لذا، فإنّ استخدام قائمة العلامات المعتادة وتوصيات **EAGLES** لن يُؤدي إلى استيعاب بعض المعلومات العربيّة ذات الصّلة، مثل جزم الفعل وتثنية العدد، التي تُشكّل جزءاً لا يتجزأ من اللغة العربيّة. وهناك جانب آخر مُهم من جوانب اللغة العربيّة وهو التوارث، حيث ترث جميع الفِئات الفرعية للكلمات أو المفردات خصائص الفئات الأم التي تمّ اشتقاقها منها.

تتضمّن هذه الورقة المباحث الرئيسيّة الآتية: وصف موجز لخصائص اللغة العربيّة؛ ثمّ مُبررات استخدام علامات الوسم المقترحة؛ ثمّ وصف لعلامات الوسم الصّرف-نحوية

العربية المقترحة؛ ثم مقارنة علامات الوَسم المقترحة في هذا البحث مع المبادئ التّوجيهية لقائمة علامات EAGLES، ووصف لأعمال أخرى مماثلة في هذا المجال. وأخيراً استعرضت الورقة نموذجاً للنّص العربي المستخدم في هذا الوسم، وقدمت قائمة كاملة لتلك العلامات مع الأمثلة.

أولاً: وصف موجز لخصائص اللغة العربيّة

العربيّة هي لُغة ساميّة، وسمتها الأساسية هي أن معظم كلماتها مبنية على جذور ويمكن تحليلها إلى جذورها. لكن يستثنى من هذه القاعدة الأسماء والحروف الشّائعة. وهناك نحو ٦٤٪ من الجذور تتكون من ثلاثة أحرف ساكنة، وهناك أيضاً جُذور تتكون من حرفين أو أربعة أو خمسة أحرف ساكنة. وتنشأ الكلمات من تلك الجذور بزيادة اللّواحق والسوابق والأحرف المزيدة. ويمكن أن تضاف السوابق واللواحق إلى الكلمات التي بُنيت من الجذور لإضافة عَدد أو نوع الجنس لهذه الكلمات. ويُسمّى هذا النظام الذي يُعنى بدراسة كيفية إنشاء الكلمات ويصف الأنماط التي تتبع لها «الصّرف» ويُطلق عليه بالإنجليزية «الصّرف الاشتقاقي» (Derivational Morphology). ورغم تشابه اللغة العربيّة مع الإنجليزية في احتوائها على المؤنث والمذكر وضمير المتكلم والغائب، لقد تمّ الاعتماد في هذه الورقة، لغرض بناء قائمة الوسم العربيّة، على نظام تدريس اللغة الذي يميز بين ثلاث حالات أو صيغ للفعل وهي (المرفوع، والمنصوب، والمجزوم)، والحالات الثلاث للاسم وهي (الرّفع، والنَصب، والجرّ).

ثانياً: مبررات استخدام علامات الوسم المقترحة

تتبع قائمة علامات الوسم الصّرف-نحوية المستخدمة هنا نظام التوسيم الذي استُخدم من جميع الطلبة الدّارسين للغة العربيّة (صغاراً وكباراً) عبر أربعة عَشَر قرناً. ويصفُ النحويون العرب اللغة العربيّة بأنها مستمدة من ثلاث أقسام رئيسيّة وهي: الاسم والفعل والحرف. وبالنّسبة لمعظم الناطقين بالعربيّة وأيضاً العديد من الدّارسين لها، تُعد هذه وسيلة طبيعية جداً لوصف اللغة. لهذا، اخترنا في هذه الدراسة فقط الفئات الرئيسيّة والفئات الفرعية. لكن بسبب الطريقة التي ترث بها الفئات الأخرى، سيكون من السّهل جداً توسيع نطاق عمل هذه العلامات لتشمل المزيد من الفئات الفرعية، أو تبسيطها وجعلها أصغر.

ويمكن وصف اللغة العربيّة باستخدام علامات وسم اللغات الهندو أوروبية، لكنها سوف تفقد كثيراً من خصائصها. على سبيل المثال، واحدة من الفئات الرّئيسية للكلام في اللغات الهندو أوروبية هي الصّفة، لكن اللغة العربيّة لا تعدّ الصّفات واحدة من الأقسام الرئيسيّة للكلام.

وجوهر الموضوع هو أن قواعد اللغة العربيّة دُرّست لقرون عدّة ، وأنّ مبادئ وصف اللغة موجودة فعلياً. وبما أن الكثير من المعرفة أصبح متاحاً اليوم وبسهولة، فمن المنطقي أن نستمد علاماتنا من هذه الثّروة المعلوماتية، وإلاّ فإنّ البديل عن ذلك هو تأسيس علامات الوسم العربيّة بناءً على علامات الوسم الهندو أوروبية، لكن القيام بذلك سوف يفقدنا كثيراً من المعلومات التي توفّرها العلامات العربيّة. أيضاً، إذا حاولنا قولبة اللغة العربيّة لتتناسب مع إحدى اللغات الهندو أوروبية، فقد نشوّه الطريقة التي يُنظر من خلالها للعربيّة من النّاطقين بها.

ثالثاً: وصف لعلامات الوسم الصّرف-نحوية العربيّة المقترحة

يناقش هذا القسم وصف علامات الوسم الصّرف-نحوية للغة العربيّة التي جمعت في هذه الورقة البَحثيّة. وتحتوي القائمة على ١٧٧ علامة وسم، و١٠٣ أسماء، و ٥٧ فعلا، و٩ أحرف، و٧ زوائد (متبقيات)، وعلامة ترقيم واحدة. وتتبع قائمة علامات الوسم نفس الخصائص الموصوفة في المبحث الثاني، وتمّ توضيحُها من خلال رسم بياني ملحق بالبحث، كما تمّ وصفها بمزيد من التفصيل من خلال الأمثلة التوضيحيّة.

وقد وضعت علامات لخمسة أقسام رئيسيّة للكلمة (هي: الاسم والفعل والحرف والزائدة وعلامة الترقيم)، ولخمسة أقسام فرعية هي: (الشائع أو العام أو المشترك، والصّحيح، والضّمير (مثل اسم الإشارة أو الأسماء الموصولة)، والعدد، والصّفة.)

ومن الأمثلة على هذه الفئات:

* الاسم النكرة المفرد والجمع المنصوب (المفعول به).
* النكرة المفرد والجمع المجرور.

* الاسم النكرة المذكّر والمؤنّث المرفوع.

ومن الأمثلة على الأسماء الموصولة: (الاسم الموصول المؤنّث والمذكّر، والاسم الموصول المفرد والجمع والمعيّن أو المحدد، ..الخ

أما الأعداد فتنقسم إلى: الأعداد الأساسية (الأصلية)، والعادية (النظامية أو الترتيبية)، والصفة العددية. ومنها الاسم النكرة المفرد والجمع ، .... الخ.

أما الأفعال، فيمكن تصنيفها إلى: الفعل التام أو الفعل الناقص أو المضارع أو الماضي أو الأمر.

في حين يمكن تقسيم الأحرف إلى: حروف جر، وظرف المكان أو الزمان، وحروف العطف، وحرف النداء، ... الخ.

رابعاً: مقارنة علامات الوسم المقترحة في هذا البحث مع المبادئ التّوجيهية لقائمة علامات EAGLES

تستند قائمة علامات **EAGLES** تقليدياً إلى اللاينية، ولا يبدو أن هناك مبرراً لاحتوائها على الفئات التي تحتوي عليها فعلياً. بالمقابل، توصف اللغة العربيّة دائماً بأنّها تحتوي على فئات معينة، وبأن الفئات الفرعية ترث خصائص الفئات الأم. وبالتالي، فإنّ علامات الوسم العربيّة يمكنها ضبط التّعميمات التي لا يمكن ضبطها في أي لغة أخرى.

وكما سبقت الإشارة، تتضمن العلامات العربيّة ثلاث فئات رئيسية (أو خمسة إذا تم تضمين علامات الترقيم والحروف الزائدة (الزوائد))، بينما تصف قائمة **EAGLES** إحدى عشرة فئة رئيسية (أو ثلاثة عشر إذا تم تضمين علامات التّرقيم والحروف الزائدة). أما الفئات الرّئيسية الإحدى عشرة التي تتضمنها قائمة علامات **EAGLES** فهي: الصّفة، والضّمير، والحرف، والظّرف، وحروف العطف، وأدوات النّداء، .. الخ.

وعلى الرغم من تشابه فئات قائمة **EAGLES** الرئيسية مع الفئات الرئيسية في اللغة العربيّة (أي الاسم والفعل)، يتم وصف الفئات الأخرى من الفئات الرئيسيّة في العلامات العربيّة على

أنها فئات أو أقسام فرعية. على سبيل المثال تُعدّ الضّمائر والأرقام من الفئات الفرعية للأسماء، في حين تعتبر حروف العطف وأدوات النّداء من الفئات الفرعية للحروف. كما أن هناك فرقا آخر بين العلامات العربيّة وعلامات EAGLES وهو العدد، حيث تُعنى الكلمات العربية بالمثنّى إضافةً للمفرد والجمع (الذي تقتصر عليه الهندو أوروبية). وهناك أيضاً الفرق المتمثّل في زمن وقوع الفعل أو صيغه، حيث يتم تعريف الأفعال في قواعد اللغة العربيّة القديمة بأنها فعل تام أو ناقص أو أمر (مجزوم). وهذا التصنيف يعد جزءاً مهماً من اللغة العربيّة، وإنّ محاولة تطويع الأفعال العربية أو تحويرها لتتناسب مع الفعل الماضي والمضارع والمستقبل التقليدية في اللغات الهندو أوروبيّة ستكون أمراً شاذاً.

الخاتمة

ناقشت هذه الورقة وصف علامات الوسم المشتقّة من قواعد اللغة العربية القديمة التي لا تتبع قواعد قائمة العلامات الهندو أوروبيّة المبنيّة على اللاتينية، إنما بُنيت على أساس تقليد السّامية لتحليل اللغة. وقد بيّنت الدراسة أنّه في العلامات العربيّة ترث كل الفئات أو الأقسام الفرعية خصائص الفئات الأم، وبالتالي يمكنها ضبط تعميمات اللغة. ولقد أوردت الدراسة ضمن ملاحقها أمثلة عمليّة تثبت أن الوسم العربي المقترح كافٍ لوصف اللغة العربيّة بشيءٍ من التّفصيل.

ICITNS 2003 International Conference on Information Technology and Natural Sciences, Amman, Jordan, pages 258-267.

Year of publication: 2003

FULL AUTOMATIC ARABIC TEXT TAGGING SYSTEM

Ghassan Kanaan, Riyad Al-Shalabi, Majdi Sawalha

بناء نظام وسم تلقائي متكامل للنصوص العربيّة

ناقش هذا البحث أهمية نظام الوسم (Tagging) للنصوص العربيّة، وقد وُجدت مثل هذه الأنظمة في معظم اللغات الغربية والآسيوية ووصل مستوى الدقة فيها ما بين ٩٥٪ إلى ٩٨٪. تكمن أهمية الذخائر اللغويّة الموسومة في عدة أوجه، أهمها: توفير معلومات قواعدية ونحوية عن النص المحدد، ومعرفة أقسام الكلام وتحديدها في الكلمات (سواء كانت أسماء وأفعالا، أو أدوات)، وإنشاء معاجم لغوية وقواعدية باستخدام بيانات لغة حقيقية. كما أنها مفيدة في التحليل الكمي للنصوص العربيّة. كان الهدف الأساسي من هذا البحث هو التعرف على الأسماء والأفعال والحروف وغيرها في نصوص اللغة العربيّة من خلال التحليل الصرفي الذي يمتاز بالدقة العالية بالإضافة إلى عنصر آخر يتعرف على الكيانات المسماة.

اعتمد الباحثون على الوثائق النصية المشكلة بالحركات وغير المشكلة لاختبار نظامهم، وحقق نتائج جيدة في مستوى الدقة، وذلك بما يقارب ٩٣٪. وبناء عليه، تبين أن المعلومات الصرفية ذات أهمية كبيرة في هذا المجال. وواجه الباحثون العديد من المشكلات أثناء دراستهم هذه، وتعود أسباب هذه المشاكل إلى طبيعة اللغة العربيّة مثل: كثرة الأسماء وتنوعها و تعدد أشكالها، كما أن بعض الكلمات تبدأ بحروف مضافة إلى الكلمة الاصليه (مثل حروف الجر وغيرها)، وأن الكلمات العربيّة غالبا ما تحتوي على اشتقاقات كثيرة لنفس الكلمة (بإضافة بعض الحروف أو حذفها)، إضافة إلى قلة عدد المعاجم والقواميس الإلكترونية. وتتسم اللغة العربيّة

بأنها لغة تشكيلية، فلفظ الكلمات العربيّة لا يمكن تحديده فقط من خلال حروف الكلمة، لكن تلعب علامات التشكيل دورا رئيسيا في عملية اللفظ إضافة إلى الهمزة والشدة التي توضع أعلى أو أسفل الحروف. واعتمد الباحثون على نهج موحد، مكون من ثلاثة أجزاء رئيسية، وهي: مدخلات ومخرجات النظام، عمليات المعالجة والمعالجات وقاعدة بيانات المشروع.

واختبر النظام المقترح باستخدام نصوص من القرآن الكريم ومجموعة أخرى مكونة من ٢٤٢ ملخصاً عربياً (غير مشكول) تم اختيارها عشوائيا من أعمال المؤتمر الوطني السعودي للحاسوب. وحسب عدد الكلمات الصحيحة، وعدد الكلمات غير الصحيحة جنبا إلى جنب مع نسبها، ثم جمعت هذه النتائج للحصول على الدقة الإجمالية للنظام، التي بلغت حوالي ٩٣٪ ونتيجة خطأ ٧٪. وصححت الأخطاء يدويا نظرا لأن مشكلة البحث تتعلق بوضع علامات على النص العربي. وقد أثرت بعض الظروف على عملية أداء هذا النظام مثل علامات التشكيل، والجذور غير الصحيحة المستخرجة عندما يتم مضاعفة بعض الحروف باستخدام حركة الشدة. وجاءت بعض الأخطاء في النظام من الكلمات المعجمية غير المدرجة في المعجم، وهذا النوع من الأخطاء يمكن تصحيحه بسهولة بإضافة هذه الكلمات إلى المعجم.

Language Resources and Evaluation, 47 (1), 33-62.

Year of publication: 2013

Supervised Collaboration for Syntactic Annotation of Quranic Arabic

Kais Dukes, Eric Atwell, Nizar Habash

الوسم النحوي يدويا للغة القرآن الكريم

يعد الترميز للغة العربيّة في القرآن الكريم موردا لغويا مهما يحتوي على أقسام الكلام المختلفة وأجزاء علم الصرف (morphology) إضافة إلى التحليل النحوي (syntactic analysis) القائم على القواعد. وقد قام مجموعة من الباحثين الراغبين في دراسة لغـة القران الكريم بإنشاء مورد لغوي موسوم يتكون من ٧٧،٤٣٠ كلمة من الذخيرة اللغويّة للقرآن الكريم (Quranic Arabic Corpus)، حيث تحتوي هذه المادة المرجعية على معلومات قواعدية ولكن بشكل غير منظّم، فعند التعامل مع النصوص الدينية يجب تحرّي الدقة الكاملة في التحليل خصوصاً عندما يتعلق بالقران الكريم. يهدف هذا المشروع إلى تقديم الوسوم الصرفية والنحوية للغة القران الكريم. وقد قام بإنشائه مجموعة مشتركة متمثلة بجامعة (ليدز) ومجموعة من الباحثين المهتمين، وقد أصبح التشارك عبر الإنترنت في الآونة الأخيرة من المهام الواسعة التي يتم استخدامها في عدة مجالات.

تهدف هذه الورقة البحثية إلى إيجاد مورد جديد يزوّد المهتمين بتحليل واف ودقيق للقران الكريم، إضافة إلى أنها تقدم نهجا جديدا متعدد المراحل مهتماً بالقضايا اللغويّة المتعلقة باللغة العربيّة. وقد تختلف هذه الدراسة من حيث البناء النحوي للجملة، حيث إن القائمين عليه قاموا بتوفير نموذج لغوي عميق قائم على أساس القواعد التقليدية التاريخية والمسماة بالإعراب. فهذه القواعد القرآنية المنسقة المألوفة تشجع علماء اللغة العربيّة والخبراء على تزويد المستخدم بتحليل شامل وصحيح للقران الكريم. لقد قام منهجهم على مجموعة مراحل مختلفة، بدأ

بوسم الكلمات التلقائي المستند للقواعد(automatic rule-based tagging)، ومن ثم مرحلة التحقق اليدوي الأولي (initial manual verification) وتليه مرحلة التدقيق المتخصص (supervised proofreading) لضمان الجودة، حيث يقوم عدد من الخبراء والمختصين بدور الإشراف والتدقيق مما يسمح لهم بمراجعة الاقتراحات التي تم الحصول عليها بواسطة المتطوعين سواء بتعديلها أو حتى الاعتراض عليها كليا، فالمهمة الأساسية التي يقوم بها المفسرون أو المتخصصون هي التدقيق الصرفي والنحوي للكلمات. وهكذا قام هذا المشروع بناء على مشاركات من بعض المتطوعين قارب عددهم مائة متطوع، حيث قدموا اقتراحات لتصويب بعض القضايا اللغوية الواردة في ترميز كلمات القران الكريم.

قام الباحثون بتقييم جودة المنهج المستخدم وكفاءته، كما قاموا بشرح التحديات التي تعترض طريق تحليل القران الكريم عبر الإنترنت، إضافة إلى مساهمتهم في وصف البرامج اللغويّة المخصصة المستخدمة للمساعدة في التحليل. وبناءً على دراساتهم تبين أن نهجهم يقدم نتائج أفضل وأكثر كفاءة من غيره من الموارد السابقة كونه يتم مراجعتها من قبل مختصين وهم يشرفون عليها بالكامل. فعلى النقيض من غيرها من الموارد، اعتمدت هذه المنهجية على دمج آليات البحث والتغذية الراجعة للحصول على مورد أكثر كفاءة ودقة، مما ساهم في إيجاد طرق جديدة لدراسة القران الكريم. ومن الجدير بالذكر أن موقعهم على الإنترنت يتلقى ما يقارب ١٥٠٠ زائر يوميا. والموقع هو: http://corpus.quran.com .

٣-٣-٣ أبحاث المعاجم الآلية

وتضم سبعة أبحاث منها ثلاثة أبحاث نوع (أ) هي : الحَاجةُ إلى قاموسٍ مَوثوقْ، ونحو تَقعيد (هَيكَلة) قامُوس عَربي- انجليزي- مَقروء آليًا باستخدام قواعد إعراب التعابير، وبناء واستخدام مورد معجمي واسع التغطية لتحسين التحليل الصرفي في العربية

وأربعة أبحاث نوع (ب) هي: بناء معجم تلقائي للغة العربية، و المحرر «ألِف» لإنشاء معاجم عربية طبيعية، و القاموس الكلاسيكي قاموس «ليمون»، و نحو نمذجة معاجم عربية متوافقة مع أطر الترميز المعجمي في لغة أنطولوجيا الويب والمنطق الوصفي.

University Bulletin – ISSUE No.16- Vol. - 63 - (4) - November - 2014.

Year of publication: 2014

The Need for a Reliable Dictionary

Sabri Elkateb

الحَاجةُ إلى قاموسٍ مَوثوقْ

المقدمة

أصبحت الحاسبات اليوم الوسيلة العملية الكفؤة الوحيدة المستخدمة في معالجة البيانات الاصطلاحية والمُعجَمية، فقُدرتها الاستيعابية التخزينية وسرعَتُها ومرونتُها اجتذبت أخصائيي المُصطلحات اللُغوية لأجلِ أتمتَة المُصطلحات، ومن ثَمّ إنشاء بنوك للمُفردات أو المُصطلحات. وفي السابق حينما لم يتوفّر نظام خاص لمُعالجة البيانات الاصطلاحية والمُعجَمية، فقد استخدمت أساليب أو منهجيات مُعالجَة البيانات القياسية.

لقد أصبحت القواميس المُحوسَبة أدوات قيّمة للغاية لإدارة موارد المعلومات، وهناك ثلاثة أنواع من القواميس: قواميس الكلمات الإلكترونية (**electronic word dictionaries**) التي تتكون من الكلمات فقط، وقواميس المفاهيم (**concept dictionaries**) التي تتميز بتصنيف التسلسل الهرمي للعلاقات(**hierarchy of relations**) ، وأخيراً موسوعات المُفردات التي تأتي على شكل شبكات دِلالية(**semantic networks**). يحتوي القاموس أيضا على هيكلية قائمة على الموسوعة المُفرداتية، إذا ما تم تضمين العلاقات الدلالية مثل، المترادفات، والمُتضادات، والعلاقات الهرمية، إلخ.

مُشكلة البحث

القاموس أو المعجم هو مجموعة من العناصر المُعجمية المُرتبة وفق ترتيب معيّن أو هيكل محدّد يتضمن المعلومات الضرورية المُتعلّقة بها. ويمكن التعبير عن هذه المعلومات بطريقة

تعريفيّة بنفس اللُغة (بحيث تُشكّل قاموساً أو موسوعة مفردات أُحادية اللُغة)، أو بِلُغة أخرى (لتُشكّل قاموساً ثنائي اللغة أو مُتعدد اللُغات). وتمثل هذه الورقة مُحاولة لاجتذاب وتشجيع مزيد من البُحوث حول الأدوات والموارد العربية أُحاديّة وثنائية ومُتعدّدة اللُغات. وتُركِز الورقة بشكل رئيسي على استغلال مزيد من السّمات المُميزة للُغة العربية ولتكريس مزيد من الاهتمام والبحث بشبكة الكلمات العربية (**Arabic WordNet**) التي ستأتي مُناقشتُها أدناه.

الحاجة لقاموس أو مُعجم موثوق

بدأت أولى محاولات حوسبة القواميس أو المَعاجم في أواخر الستينات، لأغراض اللغة والاستكشافات الحاسوبية. لكن الموارد التي كانت متاحة في ذلك الوقت لم تتمكن من التعامل مع القدر الهائل من المواد اللُغوية التي كانت بحاجة لاستكشاف. (من الأمثلة على تلك الحقبة قاموس ويبستر كوليجييت الجديد).

غير أنّ التقدم السريع في تكنولوجيا الحاسوب شَجّع على تحقيق المزيد من الطموحات للتوصل لنتائج أفضل. كما شجعت النسخ الإلكترونية من القواميس المطبوعة، والمعروفة بـاسم «القواميس القابلة للقراءة الآلية»، الباحثين على استغلال البيانات الجاهزة. وعلى الرغم من المشاكل التقنية المتعلقة بحقيقة أن تلك القواميس مكتوبة بطريقة خاصة باستخدام رموز وشيفرات يمكن استيعابها من الإنسان وليس الآلة، إلاّ أن عددا كبيرا من المشاريع كان اعتماده على القواميس الورقية (من الأمثلة على ذلك قاموس لونجمان للغة الإنجليزية المعاصرة). ورغم وجهات النظر المتشائمة التي حملها بعض الباحثين تجاه « القواميس القابلة للقراءة الآلية» مورداً مفيداً للمعلومات في سياق مُعالجة اللُغة الطبيعية، إلاّ أنّ حماسهم كان مُحفِزُه إيمانهم بأن استخراج المعلومات من تلك القواميس كان أمراً سهلاً. ولذلك شجعت نظرتهم الباحثين الآخرين، وكان من نتائج ذلك أن تمكّنت شركة ميكروسوفت من تطوير قاموس مُؤتمت بالكامل على شكل قاعدة معارف معجمية ضخمة جداً والتي عُرفت باسم "**MindNet**".

شبكة الكلمات(**WordNet**)، مورداً مُعجمياً مفاهيمياً (**Conceptual Lexical Resource**)

تعد المفاهيم هي الوحدات التنظيمية (organizational units) في نموذج شبكة الكلمات WordNet ، وهي تمثل أكثر من كلمة واحدة، لأنها تشمل: المركبات (compounds) والتنسيقات (collocations) والعبارات الاصطلاحية (idiomatic phrases) والأفعال بصيغة عبارات (phrasal verbs)، التي توسّع من نطاق فكرة تخزين الكلمات في المعجم لتشمل تخزين المعلومات المفاهيميّة (conceptual information) التي قد لا يكون لها تمثيل معجمي باستخدام كلمة واحدة». والشيء الوحيد الذي لا يفعله النموذج هو توفير تنظيم أو هيكل موضعي (topical organization of the lexicon) أو محلّي للمُعجم.

وتعدّ شبكة الكلمات WordNet قاعدة بيانات معجمية بالإنجليزية متاحة على نطاق واسع وأداة مرجعية قيّمة لهندسة اللغة والبحوث اللغوية الحِسابية. وقد فتحت الشبكة رؤى جديدة للُّغويين والمعجميين وأطلقت حقبة جديدة من أدوات اللُغة المُحوسَبة وموارد الدلالات المُعجميّة.

علاقة الكَلمة بالمعنى في شَبكة الكلمات WordNet

المُرادف هو علاقة دلالية بين كلِمتيْنِ لهما أشكال مختلفة ومعاني متماثلة، والطريقة التقليدية لتحديد المرادف هي الاستبدال. وبالإضافة إلى المرادفات، هناك العديد من العلاقات الدلالية التي حُددت بوضوح في شَبكة الكلمات WordNet، وهي: علاقة العام-الخاص (الجناس) (Hypernyms-Hyponyms)، مثل: الببغاء اسم جناس لاسم عام (الطيور)، وعلاقة الجزء بالكل(meronym-holonym)، وغيرها من العلاقات. وتشترك الأسماء في علاقة الترادف والتضاد، والعام بالخاص، والجزء بالكل، في حين ترتبط الأفعال بعلاقات المُترادفات والأضداد، بالإضافة لعلاقات العام بالخاص (الجناس) والتلازم. أما الصفات والظروف فترتبط فيما بينها بعلاقات المُترادفات والأضداد.

الحاجة إلى قواميس ثُنائية اللُغة

اللغة العربية هي اللغة الرسمية لمئات الملايين من العرب، وهي اللغة الدينية للمسلمين

جميعهم من مختلف الأعراق في جميع أنحاء العالم. لكن من المثير للدهشة، أن هناك القليل الذي تم القيام به في مجال حوسبة اللغة والموارد المُعجميّة، مقارنة مع الإطار الواسع للموارد المفاهيميّة (**lexical resources**). وقد جرى إدخال مشروع شبكة الكلمات العربية **Arabic WordNet** في رسالة دكتوراه (**Elkateb** (٢٠٠٥)) كشفت ثراء اللغة العربية، وقدرتها على الوقوف بمُفردها أو دمجها في مورد ثُنائي اللُغة أو مُتعدد اللغات مماثل لحجم وتصميم شبكة الكلمات الإنجليزية **WordNet**.

الاستفادة من السّمات الخاصّة للعربيّة

عند التعامل مع لُغات أكثر ارتباطاً باللغة الإنجليزية، يمكن تطوير شبكة كلمات متعددة اللُغات ببساطة، أما اللُغة العَربية فتمتاز بنظام واسع من الصّرف الاشتقاقي (**derivational morphology**) الذي يُجسِد العلاقات الدِلاليّة المهمة، التي يجب أن تنعكس في أي قاموس مفاهيمي. فالجذر العربي هو نوع من المفاهيم الرئيسية لا يمثّل كلمة ولكنه هيكل منه تُشتق الكلمات. وفي شبكة الكلمات العربية–الإنجليزية ثُنائية اللغة، يجب تخزين جذر اشتقاق وشكل كل كلمة في المُحتوى، لأن هذه الطريقة، التي تربط الكلمات ربطاً دِلالياً، هي أساس التوقُع للمُتحدث بالعربية. بالإضافة لذلك، يمكن للأنماط توفير السّمات المتنوعة المختلفة للنظام.

الشبكة العربية للكلمات Arabic WordNet

لقد أنشئت الشّبكة العربية وفقاً للأساليب التي طُوّرت بموجبها الشبكة الأوربية للكلمات (**EuroWordNet**)، وبهدف تحقيق أقصى قدر من التوافق عبر شبكات الكلمات، وقد جرى التركيز على الترميز اليدوي لمعظم المفاهيم الأكثر أهميةً وتعقيداً. وقد رُمزت المفاهيم والعلاقات ذات الخُصوصية اللغوية حسب الحاجة أو المطلوب.

ويواجه بناء الشبكة العربية تحدّيات لا تواجهها مثيلاتها من الشّبكات القائمة، وتشمل هذه التحديات النصوص أو الكتابة من جهة، والخصائص الصّرفية للُغات السَاميّة التي

تتمحور حول جذور المُفردات من جهة أخرى، وقد وضعت الأسس اللازمة لمواجهة تلك التحديات، فالابتكار بما يحققه من نتائج مهمة وحيوية لتطوير شبكات الكلمات هو البديل المقترح لاستبدال شبكة الكلمات الإنجليزية English بـالأنطولوجيا المقترحة المدمجة العليا(Suggested Upper Merged Ontology (SUMO)) والأنطولوجيات المماثلة لها بنفس المجال، وباتت تشكل أضخم أنطولوجيا اصطلاحية.

وعليه، وفقاً لأهداف المشروع، أصبحت الشبكة العربية (Arabic WordNet) تضُم حالياً ١١٢٧٠ قائمة مرادفات (منها (٧٩٦١ اسمية، ٢٥٣٨ فعلية، ٦٦١ نعتية (صفات)، و ١١٠ ظرفية)، وتحتوي على ٢٣٤٩٦ تعبيراً عربياً. ويشمل هذا العدد ١١٤٢ مجموعة أو قائمة مرادفات تتوافق مع الهيئات (الكيانات) المسماة التي استُخرجت تلقائياً ودُققت بواسطة المُعجَميّن.

وقد ركز أحد البُحوث الواعدة على التوسع شبه التلقائي لشبكة (Arabic WordNet) العربية باستخدام القواعد المعجمية والصّرفية. وقد كانت هناك دائماً حاجة لتوسيع نطاق تغطية الشبكة من خلال الاستفادة من مجموعة محدودة من القواعد الصّرفية العربية ذات الإنتاجية العالية لاستخلاص مجموعة من أشكال الكلمات ذات العلاقة الدِلالية من مختلف تصاريف الأفعال. وقد نُفِّذت بحوث حديثة حول الشبكة العربية باستخدام تطبيق طريقة الفهرسة الدلالية للوثائق والاستعلام لاسترجاع المعلومات باستخدام الشبكة العربية كمورد دِلالي لاستقصاء أثر الانتقال من الفهرسة على أساس الكلمة المفردة إلى الفهرسة على أساس المفاهيم.

الاستنتاجات والتوصيات

من المتوقع أن تقبل أي قاعدة بيانات مُعجمية أو اصطلاحية مجموعة متنوعة من الاستعلامات، فهناك استعلامات حول عناصر مُفردة من البيانات مثل الفئة النّحوية أو المرادف أو المتضاد أو الجناس أو تعريف المصطلح أو غيرها، كما يمكن أيضا إجراء عمليات البحث أو

الاستعلام حول مصطلح أو مصطلحات غير معروفة، في حين تكون المعلومات المتاحة حول معناها معروفة. وتوفّر شبكات الكلمات موارد للبيانات وإطاراً واضحاً للمعلومات المُعجمية، بالإضافة إلى أهميتها كموارد للكثير من التطبيقات ضمن إطار التقنيات اللُغوية. لقد أُنشِئت شبكات الكلمات في العديد من اللغات، الأمر الذي ساعد على نشر قواسمها المُعجمية المشتركة وتنوعها. إلاّ أن التّحدي القادم هو أن تصبح هذه الشّبكات مُتعددة اللُغات قابلة للتّشغيل البيني (المتبادل) (**fully interoperable**) بشكل كامل.

International Journal of Computational Linguistics Research Volume 5 Number 1 March 2014

Year of publication: 2014

Towards Structuring an Arabic-English Machine-Readable Dictionary Using Parsing Expression Grammars

Diaa Mohamed Fayed, Aly Aly Fahmy, Mohsen Abdelrazek Rashwan, Wafaa Kamel Fayed

نحو تَقعيد (هَيكَلة) قامُوس عَربي- انجليزي- مَقروء آليًا باستخدام قواعد إعراب التعابير

المُقدّمة

القواميس هي موارد غنية بالمعلومات المُعجميّة (المُفرَدة) (lexical information) ذات العلاقة بمَعاني الكَلِمات المَطلوبة للعَديد من تطبيقات تقنيات اللُغة البشرية. ومع ذلك، جَرت العادة أن يقوم الناشرون بإعداد القواميس المَطبوعة لغرض الاستخدام البشري وليس للمعالجة الآلية. أما القواميس المحوسبة فقد صممت لتكون منطقياً سهلة التعامل أو التحويل الى قواعد بيانات مُعجَميّة أو مُفردة، كما أن لها العديد من التطبيقات مثل: الترجمة الآلية والقدرة على الفهم أو الاستيعاب أو أدلة النُطق للتعرف على الكلام، أو غيرها.

وتَعرِض هذه الورقة طَريقةً لتقعيد أو هيكلة جُزئية لنُسخة قابلة-للقِراءة آلياً من قاموس المَوْرِد العربي- الانجليزي. حيث قامت هذه الطريقة بتحويل إدخالات قاموس المَورِد من سيل (تيار) من الكلمات وعلامات الترقيم إلى هياكل هَرَميّة(hierarchical structures).

ويُعدّ الهيكل الهَرمي (أو ما يُعرف بالهيكل الشجرة) عن مكونات كل إدخال من القاموس بشكل صريح. وتمثل مكونات الإدخال إدخالات فرعية، في حين يتكون كل إدخال فرعي

من عبارات تعريفية(defining phrases)، و عناوين الباب(domain labels) او الموضوع، بالإضافة إلى الإشارات المَرجِعيّة(cross-references)، و مُكافِئات التّرجمة (translation equivalences). لقد قُمنا بتَصميم الطّريقة المقترحة كخُطوات مُتسلسِلة؛ حيث يمثّل التحليل اللغوي الخطوة الرئيسية. ثم قمنا بتطبيق المُحلّل من خلال شكليّة تحليل الصّيغة التّعبيريّة النَحويّة.

مُشكلة البحث

على الرغم من أن القواميس العربية لا تحتوي على مُعايرة للهياكل الجُزئية (صُغرى) (microstructure standardization)، فقد أظهرت نتائج هذه الدراسة أنه من المُمكن هيكلتَها بشكل تلقائي أو شِبه تِلقائي بدقّة معقولة(plausible accuracy) بعد تحفيز الهيكلة الجُزئيّة فيها.

أسلوب الدّراسة

إلى جانب الاستعراض المَرجَعي للدّراسات السّابقة ذات العلاقة بموضوعها، تضمنت هذه الوَرقة المَباحث الرئيسية التالية: (١) شرح تفصيلي لتركيبة أو هيكلية قاموس المَورِد وأنواع المَعلومات الواردة في تعريفاته؛ ثمّ (٢) التّعريف الاصْطِلاحي لـ شكليّة تحليل الصّيغة التعبيرية النّحوية (parsing expression grammar formalism)؛ ثمّ (٣) خُطوات طريقة التّقعيد أو الهيكلة وأمثلة ايضاحية حولها؛ وأخيراً (٤) الاستنتاج والتّوصيات المُستقبلية.

أولاً: تركيبة (هيكلية) قاموس المَوْرد

قاموس المورد (العربي- الانجليزي) هو قاموس عام متوسط-الحجم يحتوي على ٣٣٤٦٣ كلمةً رئيسية بارزة (مُتميزة) وحوالي ٥٨٥٤٧ إدخالاً فرعيّاً. ويحتاج فهم واستكشاف تركيبة أو هيكلية القاموس فهم واستيعاب هيكليته الكُبرى وهيكليته الصُغرى وكذلك انواع المعلومات التي تحتويها معانيه أو تعريفاته.

ولقد قام مُؤلّف قاموس المورد بترتيبه حسب الأحرف الهِجائية الأولى للكلمات الرئيسية، بحيث يبدأ الإدخال بكلمة رئيسية مطبوعة بِخَطِ عريض. وفي حالة ما إذا كانت الكلمة أو المُفردة الرئيسيّة لها أكثر من معنى، يشغل كل معنى محل إدخال فرعي في سطر مُستقل. أما في حالة ما إذا كان الإدخال الفرعي يُعبّر عن نَسَق مُعيّن أو مَجاز أو أمثلة فيُشار إليها لاحقاً. ولغرض تمييز الكلمات الرئيسية فُصل كل نص إدخال فرعي باستثناء الأول عن الكلمة الرئيسية بمقدار مَسافتين. ويتكون القسم العَربي من ثلاثة حُقول: العُنوان الرئيسي (الّترويسة) والشرح والإسناد التّرافُقي، حيث يُعدّ القسم الأول إلزامياً، في حين يُعدّ القِسمان الثاني والثالث اختياريَين. ولدى الأقسام العربية الثلاثة من القاموس نفس التركيبة أو الهيكلة، حيث يَشتمل كل حقل منها على كلمة واحدة أو أكثر، بحيث يفصل بينهما فاصلة عربية أو حَرف عَطف.

في حين يتكوّن القسم الانجليزي من واحدة أو أكثر من مجموعات الترجمة المُكافئة المفصولة بواسطة فارزة (فاصلة) مَنقوطة، وكل مجموعة ترجمة مُكافِئة مُكونة من عبارة أو أكثر مفصولة عن بعضها بفاصلة عادية.

من الجدير بالذّكر أن القواميس العربية لديها مشاكل مع بعض الجوانب، يجب أن يُعالجَها مُؤلفو المَعاجم. وتُشكّل هذه المشاكل تحديات في طريق حَوسبة القواميس العربية. وفيما يلي نذكر بعض أوجُه القُصور في قاموس المورد التي من شأنها أن تُعقد مُهمّة الهَيكلة أو الحَوسبة:

أولاً: التناقض أو عدم المُطابقة.

ثانياً: الغُموض

ثالثاً: الأقواس (عَلامات الحَصر)

رابعاً: العِبارات الرئيسيّة غير المُحدّدة (غير الواضِحة)

ثانياً: تَحليل الصّيغة التّعبيرية النّحَوية (Parsing expression grammars (PEGs))

بُنيت شَكلية تحليل الصّيغة التّعبيرية النّحَوية على نماذج (Backus-Naur Forms BNFs) ودِلالات التعابير النّمطية (المُنتظمة) (Regular Expressions (REs) notations) لتشكيل

فئة بديلة من قواعد اللُغة الاصطلاحية أو المنهجية. ويعدّ نموذج (PEG) تحليلياً وليس اشتقاقياً أو مُحدِثاً، حيث اعتمدت هذه الدراسة على مكتبة بايبارسينغ (Pyparsing) والتي هي عبارة عن تطبيق لتحليل الصّيغة التّعبيرية النحويّة.

ثالثاً: طريقة أو منهجية الهيكلة (التقعيد)

من خلال استعراض عيّنات من تعاريف قاموس المَورد، وجدنا بأن علامات التّرقيم الصُغرى في قاموس المورد تحتوي على نسبة عالية من التّناسق أو الانتظام الذي يُمكِنُنا ضَبطَه وتشكيلَه. غير أنّ المعلومات الصّرفية والنّحويّة والدِلاليّة كانت مُتناثرة العبارات وغير مُتّسقة. ومن هنا، اعتمدت استراتيجيّتنا في ضبط مُدخلات قاموس المَورد على التقاط علامات التّرقيم من خلال تقنية التّحليل أولاً، ثم استخراج المعلومات المتناثرة بواسطة تقنية استخراج المعلومات.

ثمّ استعرضت الدراسة خُطوات الهَيكلة أو التّقعيد المُشار اليها أعلاه، وهي كالآتي:

١- الحُصول على بيانات القاموس.

٢- تعيين الكلمات الرئيسية للبَيانات الوَصفيّة.

٣- المُعالجة المُسبَقة (التجهيز).

٤- التّحليل اللُغوي (الإعراب).

٥- الإسناد التّرافُقي (المَرجِع المُتقاطع).

٦- التمييز أو التفريق بين المعاني.

الاستنتاجات والتّوصيات

اختارت هذه الدراسة حرف «العَيْن» من قاموس المَوْرِد لتقييم خُطوات الهَيكلة. ويُشكل حرف العين حوالي ٦,٤٪ من حَجم بيانات المَوْرِد. ةثد كان مِعيار اختيار حرف (ع) هو أن

حجمَه مُناسِب لغرض تقييم التّجربة ولاعتبارات حُقوق التأليف والنّشر. وأظهرت نتائج البحث مصادر الأخطاء وعددها وأنواعها حسب كل مصدر. ومن خلال تحليل الأخطاء حسب كل مصدر من مصادر الأخطاء، استنتجت الدراسة بأن أخطاء الإسناد التّرافُقي (المَرجِع المُتقاطع)(crossrefs) ومعنى (الفعل- يكون) كانت عبارة عن أخطاء مُكرّرة عن أول مصدرين للأخطاء: الأول (داخل-القاموس) والثاني (إدخال-البيانات) وهما مصدران معتمدان على بعضهما. وفي حال ما إذا حسبنا أخطاء مصادر الخطأ المُستقلة، تكون نسبة الدّقة حوالي ٩١٪. أما إذا أخذنا جميع الأخطاء من جميع مصادر الخطأ، فتكون نسبة الدقة حوالي ٨٧٪.

وقد أثبتت نسبة الدقة في خُطوات الهيكلة، وخاصة المُحلّل اللُغوي، الفرضيّة التي اعتمدت عليها هذه الدراسة وهي: أنّ قاموس المَورد يحتوي على نسبة عالية من التّناسق أو الانتظام في علامات التّرقيم الصُغرى بحيث يُمكِنُنا ضَبطَه وتشكيلَه تلقائياً. يُضاف لما سبق أنّ استعراض الذخائر اللغوية الثنائية للعبارات التّعريفية أظهر امتلاكها معلومات غنية مُحتملة يمكن استخلاصها وضَبطها. من جهة أخرى، فإنّ الدّرس الرئيسي المُستفاد من هذه الورقة البحثية هو أنه يُمكننا ضبط أو تشكيل القواميس العربية – التي يُعتقد أنها لا تتضمن تنسيقات مُوحدة – بدرجة معقولة من الدّقة.

وختاماً أوصت الدراسة بمتابعة البحث في المهام التالية:

١- شرح تعريفات القاموس من خلال علامات أقسام الكلام والعلامات الدلالية.

٢- تحليل وهيكلة العبارات الموضوعة بين الأقواس.

٣- حل إشارات الإسناد الترافقي بشكل تام بحيث يشير كل مرجع متقاطع (cross-references) إلى المعنى الدقيق لكلمة رئيسية مُحددة.

٤- استخراج المعلومات النّحوية والدّلالية من متن العبارات التعريفية.

٥- إزالة غموض كلمات العبارات المُحددة والعبارات المُكافئة للترجمة.

٦- تحليل حرف العطف (أو) في اللُغتين العربية و الإنجليزية أينما ورد.

Language Resource and Evaluation Conference LREC 2010 17, May 2010 - 23 May 2010, Valleta, Malta. European Language Resources Association (ELRA) 287 282-,

Year of publication: 2010

Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic

Majdi Sawalha, Eric Atwell

بناء مورد معجمي واسع التغطية واستخدامه لتحسين التحليل الصرفي في العربية

الملخص

موارد اللغة الواسعة التغطية التي توفر معرفة لغوية يجب أن تحسن الدقة والأداء لتطبيقات معالجة اللغات الطبيعية. تم بناء مورد معجمي واسع التغطية لتحسين دقة المحللات الصرفية ومحللات أقسام الكلام في النص العربي. حيث تم جمع ٢٣ معجم محوسب متوفرة مجاناً على الإنترنت، وجمعها جميعاً في مورد معجمي واحد من خلال استخراج المعلومات من تنسيقات مختلفة ودمج المعاجم العربية التقليدية. تم تقييم المورد المعجمي الواسع التغطية من خلال حساب شموله لكلمات القرآن الكريم، وموارد أخرى عبر الإنترنت.

المقدمة

الصناعة المعجمية هي الجزء التطبيقي من علم المعاجم ويهتم بفرز المدخلات والاشتقاقات ومعانيها وترتيبها بهدف بناء المعجم. وقد ساهم التطور التكنولوجي واستخدام الحاسوب بتسريع مشاريع بناء المعاجم، ويشار إلى استخدام الحاسوب في بناء المعاجم بالحوسبة المعجمية (computational lexicography). وعليه، خزنت قواميس اللغة الإنجليزية الحديثة

باستخدام قواعد بيانات معجمية محوسبة بواسطة لغة برمجة (SGML) التي تعتمد XML وأي نظام إدارة قواعد بيانات (DBMS) وقواعد البيانات العلائقية.

أما المعاجم العربية التقليدية فهي غير متوفرة في قواعد بيانات معجمية محوسبة، وبأي حال فهي تتبع طرق ترتيب تختلف عن المعاجم الإنجليزية الحديثة. والمعاجم الإنجليزية ترتب الكلمات أبجدياً متبوعة بمعنى الكلمة، بينما المعاجم العربية رتبت على أساس الجذر كمدخلات أساسية، ويتبع الجذر تعريفاً من الممكن أن يكون على صفحات بحيث يحتوي على جميع الاشتقاقات الممكنة لهذا الجذر ولا تتبع لتنسيق خاص أو ترتيب معين.

الصناعة المعجمية التقليدية

الصناعة المعجمية أحد أبرز وأعمق علوم اللغة العربية، فكان أول المعاجم العربية كتاب العين للفراهيدي، وبنيت معاجم أخرى كثيرة اختلفت في حجمها أو ترتيبها أو الهدف من بنائها.

خطوات معالجة المعجم العربي

تم جمع ٢٣ معجماً عربياً من موارد مختلفة على الشابكة، ومكتبة المشكاة الإسلامية التي توفر معظمها بصيغة محرر النصوص (MS Word)، ولكن لكل منها طريقة ترتيب معينة تختلف عن الآخر.

فبعد تحويلها يدوياً إلى تنسيق موحدة من خلال اختيار التنسيق الأكثر شيوعاً للجذور، تم استخراج الكلمات والمعاني آلياً باستخدام برامج متخصصة، تخزن النتائج بعدها في قواميس منفصلة تضم الجذور والكلمات ومعانيها، ثم جُمعت معلومات المعجم باستخدام خوارزمي تجميع ضمن مورد معجمي واسع التغطية واحد.

الخطوة الأولى بتحويل المعاجم من صيغة doc أو HTML إلى نص قياسي بالصيغة الموحدة (Unicode UTF-8)، ثم تحصى تكرارات الكلمات وحجم المفردات للنصوص المشكولة وغير المشكولة لكل معجم.

تحليل نصوص المعاجم منفصلة:

وكل معجم بني بطريقة مختلفة في ترتيب الجذور والمدخلات. كتبت المعاجم العربية بطريقة مقروءة آلياً دون استخدام أي تمثيل محوسب ، لذا عولج كل معجم منفصلاً عن الآخر، باستخدام برامج متخصصة.

الخطوة الهامة التي تسبق المعالجة هي التنسيق اليدوي لكل نص من نصوص المعجم إلى تنسيق موحدة باختيار التنسيق الشائع الجذور جميعها في المعجم.

البنية الشائعة لهذه المعاجم هي بنية تعتمد على الجذر والتعريف الذي يضم كافة الكلمات المشتقة مع معانيها، ثم يتم استخراج الجذور والكلمات المشتقة باستخدام برنامج خاص.

ثمّ رُبطت مدخلات المعجم مع معانيها دون طريقة محددة لترتيبها. تأتي المدخلات مرتبطة بجميع أشكال اللواصق والإضافات كأدوات الربط والضمائر. ولأن اللواصق تشكل تحدياً لبناء المورد المعجمي واسع التغطية، فقد استُخدام محلل صرفي لفصل المدخلات من لواصق الكلمات بمختلف صورها.

هنا تأتي العديد من الكلمات في قسم التعريف لا علاقة لها بالجذر، لذا تأتي مرحلة التحسين للتأكد من أن أزواج الكلمات والجذور تتناسب مع الاشتقاقات المعروفة للجذر، من خلال الخطوات الآتية:

١- فحص الحروف: إذا كان جميع الحروف المكونة للجذر موجودة في الكلمة المحللة. ننتقل إلى الخطوة الثانية

٢- ترتيب الحروف: إذا كانت جميع الحروف في الكلمة المحللة تظهر بنفس ترتيب حروف الجذر، لذا من الممكن أن تكون الكلمة صحيحة.

تجميع المعاجم المعالجة إلى مورد لغوي واسع:

بعد تحليل كل معجم نستخدم خوارزمية دمج لبناء معجم واسع، الخوارزمية تبدأ باختيار

معجم لسان العرب كنواة للمورد المعجمي الواسع التغطية، ثم تأتي بقية المعاجم تباعاً مع نسبة السجلات الموجودة فيها.

التقييم

عملية التقييم أشارت إلى تغطية المورد المعجمي الواسع لأنواع مختلفة للنصوص، القرآن الكريم، الذخيرة العربية الإلكترونية، والذخيرة العربية المعاصرة، التي حوسبت بطريقتين: الأولى – مقابلة دقيقة للكلمات غير المشكولة بنسبة تتراوح ٦٥٪-٥٧٪.

الكلمات العربية في أي نص تأتي بأشكال مختلفة للواصق التي تتصل بها، لتجعل من مهمة مقابلة الكلمة مع مدخلات المعجم مهمة غير سهلة وتقلل من شمولية المورد المعجمي الواسع.

ولحل هذه المشكلة تم بناء مجذع (Lemmatizer) لمعالجة بيانات كبيرة وحقيقية، يعتمد فيها على المورد المعجمي الواسع لاستخراج جذع وجذر الكلمة، فإذا وجد الجذع (الجذر مع اللواصق والإضافات) في المعجم يتم استرجاعه وإضافته إلى المورد المعجمي الواسع.

نسبة الشمول سجلت كالآتي:

١- ٨٥٪ من الكلمات مع الكلمات الوظيفية.

٢- ٨٢٪ من الكلمات دون الكلمات الوظيفية.

الخلاصة

شاهدنا خلال هذه الدراسة عملية بناء مورد معجمي واسع للعربية باستخدام تطبيقات معالجة اللغة الطبيعية (NLP)، حيث طور المورد من خلال دمج موارد لغوية مختلفة.

وكانت خطوات العمل بتحليل المعاجم أولاً، ثم تحويلها إلى تنسيق موحد، ثم استخراج الجذر مع الكلمات المشتقة منه، ثم دمج المخرجات من الخطوة السابقة إلى مورد معجمي واسع، وقُييم المورد المعجمي الواسع التغطية وذلك بحساب شموليته باستخدام طريقتين: مقابلة الكلمات مع كلمات المعجم وسجلت نسبة ٦٧٪، والثانية باستخدام المجذع وسجل ٨٢٪.

International Journal of Computing & Information Sciences, Vol. 2, No. 2, Augus t 2004.

Year of publication: 2004

Constructing an Automatic Lexicon for Arabic Language

Ghassan Kanaan, Riyad Al-Shalabi

بناء معجم تلقائي للغة العربية

تعد المعاجم أحد أهم الموارد في مجال اللغويات التطبيقية (applied linguistics) المعنية ببناء المفردات واستخدامها، حيث يمكن أن تكون هذه المعاجم ورقية أو على شكل برامج الكترونية محوسبة ومتوفرة على شبكة الإنترنت، تعنى المعاجم بتوفير بعض المعلومات الهامة التي تخص المفردات مثل بعض التعابير والأمثال ودلالات الكلمات إضافة إلى المعنى، ويكون المعجم مجموعة من التمثيلات للكلمات المستخدمة، وقد تحتوي هذه التمثيلات على معلومات تتعلق بتشكيل الكلمات، علم الأصوات(phonology) ، البناء النحوي syntactic structure)) إضافة إلى دلالات الكلمة (semantics).

كان الهدف الرئيس من القيام بهذه الدراسة هو تصميم وتنفيذ نظام لبناء المعجم التلقائي (Automatic Lexicon) للغة العربية، بحيث يشتمل على معلومات محددة مثل جذور الكلمات وأنماطها و أجزائها والمقاطع التي يمكن إضافتها للكلمة ونوعها (اسم، فعل، صفة، ظرف أو غيرها) أو السمات المعجمية مثل الأعداد والجنس والحالة والتعريفات وغيرها. لذلك كانت هناك حاجة ماسة إلى وجود معجم جيد للعديد من تطبيقات اللغات الطبيعية (Natural Language applications) مثل: التحليل وتحديد الجمل الاسمية والجمل الفعلية وغيرها. استخدم الباحثون في نظامهم العديد من القواعد التي تستند إلى قواعد اللغة العربية لتحديد أقسام الكلام (نوع الكلمة) والسمات المعجمية ذات الصلة بالكلمة. فقد قام الباحثان باختبار

النظام بناء على استخدام وثائق نصية عربية مشكلة وغير مشكلة (vowelized and non-vowelized Arabic text documents) مأخوذة من القرآن الكريم إضافة إلى ٢٤٢ ملخصاً عربياً اختيرت عشوائياً من أعمال المؤتمر الوطني للحاسوب في المملكة العربية السعودية، وقد حقق هذا النظام مستوى جيد من الدقة بما يعادل ٩٦٪ تقريبا. وقد نوقشت العوامل الكامنة وراء بعض الأخطاء وكيفية تعزيز معدل الدقة ليصل إلى مستويات أعلى.

بعد قيام الباحثين ببعض التجارب والاختبارات تبين لهما أنه عند حساب كفاءة النظام يجب تجاهل الأخطاء المتعلقة بجذر الكلمة وتحليل جميع الأجزاء الأخرى؛ لأن تركيز النظام يقوم على بناء معجم عربي تلقائي، فهذه الأخطاء تنتج عن وجود بعض الظروف التي يصعب السيطرة عليها. فالجذوع مبنية على الجذر الثلاثي للكلمة، لكن هناك بعض الكلمات التي لا يوجد لها جذور ثلاثية (رباعية أو غيرها)، فمثل هذه الكلمات غير مشمولة في النظام المقترح. ومن العوامل الأخرى التي تؤثر سلباً على أداء وكفاءة هذا النظام جذور الكلمات التي تحتوي على حرف مضعف أو مشدد (الشدة) وهي ليست علامة تشكيل، وإنما تعني أن الحرف مضعف عند اللفظ. كما أن بعض الأسماء الدالة على الجمع تتشكل بطريقه شاذة (جمع التكسير) (irregular plural) أو أحيانا بعض كلمات المفرد أو المثنى تبدو وكأنها جمع، وهذا أيضاً يؤثر على أداء النظام.

8th International Conference on Information and Communication Systems (ICICS), pp. 70-75. IEEE

Year of publication: 2017

ALIF editor for generating Arabic normalized lexicons

Samia Ben Ismail, Hajer Maraoui, Kais Haddar, Laurent Romary

المحرر "أَلِف" لإنشاء معاجم عربية طبيعية

أصبح توحيد المعلومات المعجمية ضرورة واضحة، خاصة بالنسبة للمجتمعات اللغوية بهدف تحقيق التوافق وإمكانية التبادل بين تطبيقات معالجة اللغات الطبيعية المختلفة. وفي الواقع، فإن المجتمع الناطق باللغة العربية المتعامل مع المعجمية يتكيف مع هذه الفكرة لإنشاء قواعد بيانات معجمية موحدة لصالح التكامل والتقييم وقابلية العمل المشترك نحو الموارد المعجمية. ومع ذلك، فإن مهمة إنشاء قواعد البيانات المعجمية هذه ليست مهمة سهلة للمستخدم اللغوي، فاللغويون لا يملكون المعرفة الخاصة بتطوير الحوسبة لإنتاج موارد معجمية موحدة على أساس المعايير القائمة. وعلاوة على ذلك، حتى لو كان اللغوي على دراية بهياكل المعايير، فإنه لا يمكنه معالجة إنشاء المعاجم إلا يدوياً، وهذه المهمة تستغرق وقتاً طويلاً. لذلك، فإن وجود محرر معجمي يستند إلى نماذج وتصاميم للمعايير يمكن أن يساهم في تبسيط وصف قواعد البيانات المعجمية، وفي الامتثال لقيود معينة خاصة عند وجود مدقق قيود. إلى جانب ذلك، يمكن أن يكون المحرر المعجمي أساساً لأنشطة التدريب القائم على الحاسوب وداعماً للمستخدمين من غير ذوي الخبرة. لذلك، فإن إنشاء هذا المحرر يعالج عدداً من المشاكل الخطيرة.

إن تطوير المعجم العربي النحوي الصرفي ليس مهمة سهلة، وتتطلب اعتماد معيار ثابت لقابلية العمل المشترك وقابلية التبادل للموارد المعجمية. ومع ذلك، فإن العمل البحثي الذي

يتناول التطبيع للموارد العربية المعجمية لم يكتمل تطوره بعد، وخاصة فيما يتعلق ببعض المعايير مثل مبادرة ترميز النص(Text Encoding Initiative (TEI)) . في هذه الورقة، يهدف الباحثون إلى إنشاء محرر معجمي (lexicon editor) للغة العربية مع مدقق القيود(constraints checker) مع مراعاة المعايير الدولية (ISO)، وإطار الترميز المعجمي(Lexical Markup Framework (LMF)) والمبادئ التوجيهية لمبادرة ترميز النص (TEI) . وقد استخدم الباحثون نهجاً لغوياً يتكون من عدة خطوات لتطوير هذا المحرر. النموذج الأولي للمحرر المسمى أَلِف(ALIF) يمكن أن يضمن بناء نوعين من الملفات المعجمية الناتجة: أحدهما في إطار الترميز المعجمي (LMF) ، والآخر في مبادرة ترميز النص(TEI) . وهكذا فإن الهدف الأساسي من هذه الورقة هو إنشاء محرر يستند على إطار الترميز المعجمي (LMF) ومبادرة ترميز النص (TEI) مخصص لإنشاء المعاجم العربية التي تشمل المستويات الصرفية والنحوية والدلالية. ولتحقيق هذا الهدف، أجرى الباحثون دراسة معمقة للكلمات العربية (الأفعال والأسماء والحروف) على مستويين: المعجمي والنحوي، إلى جانب الوصول إلى توحيد للخزينة المعجمية. كما استند الباحثون على معيار الأيزو (ISO) ٢٤٦١٣ وإطار الترميز المعجمي (LMF) والمبادئ التوجيهية لمبادرة ترميز النص (TEI).

يستند تقييم هذا النظام على قاعدة بيانات معجمية تحتوي على جميع الأشكال المشتقة الناتجة عن معجم يحتوي على عشرة آلاف فعل أساسي. وتظهر القيم التي حُصل عليها من القياسات أن المحرر يعطي نتائج مشجعة. ويمكن استخدام هذه النتائج في مستويات أخرى من التحليلات. وعلاوة على ذلك، فإن وقت التنفيذ هو الوقت الفعلي الصحيح إذ يحتاج إنتاج الملفات المعجمية إلى بضع دقائق فقط.

4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 2016, pp. 325-330

Year of publication: 2016

Classical Dictionary Al-Qamus in lemon

Mustapha Khalfi, Ouafae Nahli, Arsalane Zarghili

القاموس الكلاسيكي قاموس "ليمون"

تزايدت الحاجة إلى إغناء المحتوى الرقمي للغة العربية الكلاسيكية في الآونة الأخيرة، ولهذا فإن هذا البحث يهدف إلى اقتراح تمثيل معجم «القاموس المحيط» ((Al-Qamus Al-Muhit (AQAM)) في شكل قياسي رقمي سمي ليمون (lemon) وهذه التسمية جاءت من LExicon Model for Ontologies))، فقاموس ليمون يُعدُّ ترميزا رقميا يستخدم الويب الدلالي لاستخراج الأوصاف اللغوية.

يعتمد مبدأ عمل هذا القاموس على مرحلتين أساسيتين، هما أولاً مرحلة الترميز والتجزئة (Encoding and segmentation)، ثانيا مرحلة الاستخراج (Extraction). في مرحلة الترميز يتم تمييز الفصول الرئيسية من المعجم والفصول الفرعية بالرموز «١» و «٢» على التوالي، ويتم وضع علامة «@» عند تغيير الجذر (Root change)، وكل سطر يحتوي على وصف معجمي يجب أن يبدأ بـ «@» أو «$» ويجب أن ينتهي بحرف. وعلاوة على ذلك، يتخلل محتوى النص المعجمي عدد من رموز الترقيم، بحيث يكون هناك تسلسل من المصطلحات المتتالية التي تسمى كلمة مجردة (lemma) ، وتكون مفصولة بنقطتين رأسيتين «:»، ثم تتبع بمجموعة من المعاني مفروزة بفواصل «;».

أما بالنسبة لمرحلة الاستخراج فإنها تشتمل على خطوتين، هما خطوة استخراج عائلة الجذر (Extraction of root family) وخطوة استخراج الكلمات المجردة (Extraction of lemmas). وفي كل خطوة تُحدد العديد من قواعد التحكم من أجل إجراء العمليات بطريقة كاملة وفعالة.

ففي خطوة استخراج عائلة الجذر يُحدد جذر الكلمة بحيث يرمز (C1) إلى الحرف الأول من جذر الكلمة ويشير (C2) إلى الحرف الأوسط. كما يشير (C3) إلى الحرف الأخير. فمثلا كلمة «مدرسة» جذرها «د» C1= «ر» = «C2» «س» = «C3» وبعد ذلك يُقارن الجذر المستخدم مع جميع الاستنباطات الموجودة المشتقة من هذا الجذر.

الخطوة الثانية من الاستخراج تهتم باستخراج الكلمات المجردة بعد التحقق من أنها تحتوي على جذر الكلمة من C1 إلى C3، وبالإضافة إلى ذلك، تُطبق العديد من القواعد أو التعبيرات العادية لتحديد أجزاء الكلام (Part Of Speech(POS)). فعلى سبيل المثال، لاستخراج الفعل تتم مقارنة الكلمات المجردة مع جميع الأشكال الفعلية المشتقة من الجذر حسب قواعد السياق التالية «فَعَلَ يَفْعَلُ فعَّلَ اِفتَعَلَ اِنفَعَلَ أفعل تَفاعَلَ تَفعيل فَعَال فَعَلَ فَعّاليّة». وهكذا بالنسبة للأسماء، إذ يجري اشتقاقها عن طريق قواعد سياقية خاصة بذلك. وفي نهاية المطاف تجري مطابقة الكلمات المجردة التي خلص إليها هذا المورد مع المصطلحات المتوفرة في نسخة الوردنت العربية (ARABIC WORDNET) وذلك للربط بين مختلف العناصر المعجمية والمعاني الصرفية للكلمات للحصول على مورد يساعد اللغة العربية على التطور في مختلف الجبهات.

Terminology and Knowledge Engineering 2014, Jun 2014, Berlin, Germany. p. 10, 2014

Year of publication: 2014

Towards modeling Arabic lexicons compliant LMF in OWL-DL

Malek Lhioui, Kais Haddar, Laurent Romary

نحو نمذجة معاجم عربية متوافقة مع أطر الترميز المعجمي في لغة أنطولوجيا الويب والمنطق الوصفي

إن مهمة تشكيل وتصميم قواعد بيانات معجمية قابلة لإعادة الاستخدام هي مهمة حاسمة من شأنها تطوير وتحسين مختلف المجالات ولاسيما معالجة اللغات الطبيعية (Natural Language Processing (NLP)) والويب الدلالي (Semantic Web). وفي هذا السياق، فإن هذا البحث يسعى إلى نمذجة إطار ترميز لغوي (Lexical Markup Framework) (LMF) ضمن لغة أنطولوجيا الويب (Web Ontology Language) والمنطق الوصفي (Description Logics (OWL-DL))، حيث تكمن أهمية هذا البحث في تركيزه على طرق القياس المرجعية المعتمدة على أطر الترميز اللغوية التي تهتم بنمذجة الهياكل المعجمية (modeling lexical structures). في البداية قدم مؤلفو هذا البحث عرضاً سريعاً لإطار الترميز اللغوي، ثم حددوا ثلاث لغات فرعية لتعريف الأنطولوجيا التي يمكن استخدامها بسهولة من مستخدمين محددين، وهي: لغة أنطولوجيا الويب الخفيفة (Lite) والكاملة (Full) والمنطق الوصفي(Description Logics (DL)) وبعد المقارنة بين هذه الأنواع الثلاثة، تم اختيار لغة أنطولوجيا الويب مع المنطق الوصفي.

النموذج المقترح في هذا البحث يتألف من سبع خطوات أساسية، وتُلخص كما يأتي:

١- بناء كيانات ذات كفاءة عالية: هذه الخطوة تنفذ من أجل تبسيط بعض الإدخالات في نمذجة لغة أنطولوجيا الويب، حيث حُددت خمسة كيانات لهذا الغرض، وسميت ب (xsd ، owl، rdf، rdfs، lmf).

٢- الفضاءات الاسمية المستخدمة (Used Namespaces): وتستخدم من أجل جعل الأنطولوجيا مفهومة وغير غامضة، حيث يُعرَّف مكون جديد يسمى الفضاء الاسمي، ويُعدُّ هذا المكون مؤشرًا على المفردات المستخدمة في الأنطولوجيا.

٣- رأس وفئات إطار الترميز اللغوي: في هذه الخطوة تُوصف مجموعة من التأكيدات بعد تعريف فضاءات الأسماء، هذه التأكيدات تزين ملف النمذجة بالتعليقات، وعلامات الترقيم، ومراقبة إصدار وإدراج المزيد من الأنطولوجيات.

٤- الفئات الفرعية لإطار الترميز اللغوي: بصفة عامة، تحتوي جميع المصطلحات على قائمة من القيود، والفئات الفرعية وهي واحدة من تلك القيود.

٥- خصائص إطار الترميز اللغوي: تُفسر بعض المعلومات العامة والمحددة على أنها سمات نموذج إطار الترميز اللغوي فعلى سبيل المثال، المعلومات العامة هي فئة إدارية عامة السمات، مثل ترميز اللغة أو ترميز النص البرمجي التي هي مناسبة لجميع الموارد المعجمية.

٦- علاقات إطار الترميز اللغوي: في هذه الخطوة تُحدد علاقات لإطار الترميز اللغوي بقائمة من قيود النطاق والقيود المشتركة، فعلى سبيل المثال: «امتلاك المعجم (Has lexicon) « هو تقييد خاص لـــ «خصائص العناصر(ObjectProperty)»

٧- عدد العناصر في مجموعة إطار الترميز اللغوي: في هذه الخطوة تُعرَّف المجموعات على شكل قيود، حيث تجعل هذه المجموعات النموذج المصمم أكثر ثراء من حيث القيود.

٨- وأخيراً فإن إيجاد مثل هذا النموذج من المعاجم في أي لغة يجعل من السهل جدًا بناء العديد من التطبيقات على هذه اللغة، وهذا بالضبط ما أُثبت من خلال تطبيق هذا النموذج.

٣-٣-٤ أبحاث الأنطولوجيا

يُعدّ علم التّوصيف (الأنطولوجيا) من بين أقوى أدوات نشر المعرفة لنمذجة وإدارة التّطبيقات المُختلفة لمعالجة اللغات الطبيعية واسترجاع المعلومات والشّبكة الدلالية. وفي هذا السياق، أصبحت الدراسات والأبحاث المتعلقة ببناء الأنطولوجيا منتشرة على نحو متزايد في مجال علوم الحاسوب.

تضم هذه المجموعة عشرة أبحاث، بينها بحثان من النوع (أ)، الأول يتعلق بمحاولة إيجاد أنطولوجيا عربيّة تُحَدِّد الأنماط الصّرفية المعجمية لاستخراج العلاقة الدلاليّة والثاني يتعلق بأداة بحث قرآنية عربية مصمّمة على أساس الأنطولوجيا. أما الأبحاث الثمانية من النوع (ب)، فشملت دراسة استقصائية لمنهجيات الأنطولوجيا الحاسوبية للتعيين الدلالي في القرآن الكريم وتطبيق لأسماء المكان في القرآن الكريم المعتمد على أنماط معاجم اللغة العربية، وتقديم البناء الآلي للأنطولوجيا باستخدام النصوص العربية، والتوجه نحو استخراج مفاهيم الأنطولوجيا العربية، واستخدام الأنطولوجيا في اللغة العربية لتحسين عمليات استرجاع إجابات الأسئلة في محركات البحث، ومشروع أنطولوجيا اللغويات العربية المسمى بالخليل، ونموذج الأنطولوجيا العربية للمحاسبة المالية ، وإطار السياق الدلالي القائم على علم الأنطولوجيا لمحتويات الويب العربية.

Proceedings of the14th International Conferenceon Enterprise Information Systems (SCOE-2012), pages342-348

Year of publication: 2012

Towards an Arabic Ontology Defining Morpho-lexical Patterns for Semantic Relation Extraction

Mohamed Mahdi Boudabous, Fatiha Sadat, Lamia Hadrich Belhuith

نحو أنطولوجيا عربيّة تُحَدِّد الأنماط الصّرفية المعجمية لاستخراج العلاقة الدلاليّة

المقدمة

اقترحت هذه الورقة البحثية طريقةً مبتكَرة لتحديد الأنماط الصّرفية المعجميّة المستخدمة للكشف عن العلاقة الدلالية بين الأسماء العربية. وتستند هذه المنهجية إلى الذخيرة اللغوية للدراسة المعتمد على موسوعة منشورة على الشّابكة. وتكون هذه الذخيرة اللغوية من مجموعة من المقالات المختارة على أساس قاعدة بيانات تحتوي على أزواج من المفردات المرتبطة بعلاقات دلالية. لقد نُفِّذت الأنماط المُحددة باستخدام «منصّة نوج" (NooJ platform) . وجاءت نتائج تقييم النّمط (pattern evaluation) مشجّعة جداً.

تناولت هذه الدراسة المباحث الرئيسية التالية: المبحث الأول: مفاهيم أساسية لبناء الأنطولوجيا؛ المبحث الثاني: لمحة عامة عن المنهجيات أو الأساليب المتبعة في استخراج العلاقات الدلالية؛ المبحث الثّالث: يستعرض أسلوب الدراسة المقترح لتحديد أو تعريف الأنماط الصّرفية المعجميّة من الموسوعة الموجودة على الشّابكة، ويُعطي تفاصيل حول المراحل المناظرة. المبحث الرابع: تناول تنفيذ الأنماط المحددة، ومناقشة النتائج التي تمّ التوصّل إليها؛ وأخيراً الاستنتاج والتّوصيات.

مُشكلة البحث

ظهرت على مدى السّنوات الماضية العديد من منهجيات بناء الأنطولوجيا. لكن اختلفت منهجيّات التعلُم وفقاً لموارد البيانات المستخدمة. ركّزت هذه الدراسة – من جهتها – على التّحليل اللغوي للذخائر اللغوية التي تسمح بالانتقال من المستوى اللغويّ إلى مستوى المفهوم. ونظراً لعدم وجود أدوات لتحليل اللغة العربية ومدى تعقيد المعالجات الخاصة بهذه اللغة، اقترحت هذه الدراسة أنماطاً صرفية معجميّة للكشف عن العلاقة الدلالية بين مفاهيم الأنطولوجيا. وتظهر أصالة أو حداثة هذا الأسلوب في تحديد الأنماط الصّرفية المعجميّة المستخدمة من موسوعة (ويكيبيديا) العربية. ومن ثَمّ فإن الأنماط المحدّدة تسمح باستخراج العلاقات الدلالية ما بين الأسماء.

المبحث الأول: المفاهيم الأساسية

إنّ مفهوم الأنطولوجيا موروث عن تقليد فلسفي يركز على علم الوجود(**science of being**). وعلى مدى العقدين الماضيين، ظهرت العديد من التعاريف وأنواع الأنطولوجيا منذ إدخال هذا المفهوم في مجال علوم الحاسوب. في هذا المبحث، نستعرض مختلف التعاريف، وعناصر علم الأنطولوجيا، وأنواع هذا العلم، ومنهجيات بناء الأنطولوجيا.

بعد بناء أوّل جيل من الأنطولوجيا، حدّد الباحثون أبعاد تصنيف الأنطولوجيا. وفي الواقع، ظهرت أبعاد مختلفة لتصنيف علم الأنطولوجيا: (الأنطولوجيا العامّة أو الشاملة، والحقلية أو المتخصّصة، وأنطولوجيا المهمات، وأخيراً أنطولوجيا التطبيق). وفي مجال هندسة علم الأنطولوجيا، تمّ اقتراح العديد من المنهجيات لبناء الأنطولوجيا. وفي الواقع، كانت أول منهجية لبناء الأنطولوجيا تُسمى «من الصفر»(**from-scratch**)، وكانت تهدف إلى تصميم عملية لبناء الأنطولوجيا في غياب المعرفة. ونظراً للقيود التي كانت تفرضها هذه المنهجية، اقترح الباحثون منهجية إعادة هندسة (**re-engineering**) الأنطولوجيا، وبعد ذلك، اقترح الباحثون منهجيّة جديدة تتبع نهجاً تعاونياً يعتمد على تدخل الناس الموجودين في أماكن مختلفة،

وأطلق عليها «منهجية البناء التّعاوني (cooperative constructing methodology).

في هذه الورقة، كان الاهتمام بأساليب التعلُم من النصوص المبنية على أساس التقنّيات اللغوية. بالإضافة إلى تلك الطرق المستخدمة في معالجة اللغات الهندية الأوروبية التي تستخدم أنماطا معجمية للتعرف على العلاقات الدلالية.

المبحث الثاني: لمحة عامّة حول منهجيات أو أساليب استخراج العلاقات الدلاليّة

يقدم هذا المبحث لمحة عامة عن التقنيات المختلفة لاستخراج العلاقات الدِلالية. وتعدّ عملية التحديد التلقائي للعلاقات الدلالية في النص مشكلة صعبة، على الرغم من أهميتها للعَديد من التطبيقات. ومن هنا، فقد اقتُرِح العَديد من أساليب استخراج العلاقات الدلالية من النصوص، التي يمكن تصنيفها إلى ثلاث فئات رئيسية: التقنيات الإحصائية؛ والتقنيات اللغوية؛ والتقنيات الهجينة.

أما التّقنيات الإحصائية فتعتمد أساساً على مبدأ اعتبار أن المفردات الموجودة معاً تربطها بقوة علاقات دلالية. في حين تعتمد التقنيات اللغوية على هيكليّة أو تركيب الجملة أو النصّ. وأخيراً، تجمع التقنيات الهجينة بين الفئتين الآنفتين، بحيث تستخدم عموماً التّوزيع الإعرابي أو النّحوي للمفردات أو التعابير لاستخراج العلاقات بينها.

المبحث الثالث: الأسلوب المقترح لاستخراج العلاقة الدلاليّة

في هذا المبحث، نقترح أسلوباً لاستخراج العلاقات الدلالية من النصوص لبناء أنطولوجيا معجمية من الموسوعة الإلكترونية على الشّابكة. واعتمد أسلوبنا على تحديد مجموعة من الأنماط الصّرفية الخاصّة بكل علاقة بحيث يمكن أن تُستخدم لاستخراج العلاقات بين المفاهيم.

وتَضمّن هذا الأسلوب المقترح مرحلتين، تمثّلت المرحلة الأولى بالمعالجة الأوّلية للمتن، وتتكوّن المرحلة الثانية من تحديد (تعريف) الأنماط والتّحقق منها. وتضمن الأنماط المحدّدة الانتقال من المستوى النّصّي إلى مستوى المفهوم وتحويل غير الاصطلاحي إلى اصطلاحي. وفيما يلي الخطوات المختلفة لهذا الأسلوب:

١- المعالجة المسبقة للذخيرة اللغوية: حيث يتم من خلال هذه الخطوة تحديد أو تعريف العلاقات الدلالية الخاصة بمجال الأنطولوجيا، ثمّ تُبنى الذخيرة اللغوية، وهو أمر حيوي وبسيط. وأخيراً يتم إجراء المعالجة المسبقة للذخيرة اللغوية (التقطيع أو التجزئة واستخلاص الجمل والتحليل الصرفي).

٢- تعريف أو تحديد الأنماط والتحقق من صلاحيّتها: في هذه الخطوة، يُحدد أو يُعرّف النمط الصّرفي المعجميّ كتركيب أو نمط لغوي مكوّن من مجموعة كلمات و/ أو أقسام صرفيّة وفق ترتيب معيّن. ثمّ يتمّ التحقق من مدى صلاحية الأنماط المحدّدة.

المبحث الرابع: تنفيذ أو تطبيق الأنماط(**patterns**)

بمجرد التحقق من صحة الأنماط من خبير في هذا المجال، تأتي الخطوة التالية الممثلة في تنفيذ جميع الأنماط المحددة. لذلك، استخدمنا منصة نوج **NooJ platform** التي سمحت لنا بتمثيل أنماط محددة لكل علاقة كقاعدة نحويّة لغويّة. وفي الواقع، جرى تمييز العناصر الصّرفية المعجمية (قبل وبعد وأثناء) بين المصطلحات ذات الصلة من خلال وضع العلامات. وتظهر هذه العناصر بين قوسين في النمط.

وجمع بين القواعد النّحوية ضمن قاعدة نحوية واحدة تُسمى «القاعدة الأساسية» من أجل تطبيق جميع قواعد اللغة على جميع الجُمل في الذخيرة اللغوية لاستخراج جميع علاقات الأنطولوجيا.

ولغرض تقييم أثر فعالية الأنماط المحددة على العلاقات الدلالية، طبقت القاعدة النّحوية الأساسية على ذخيرة الاختبار التي تتكون من ٣٠٠ مقالة مستخرجة من موسوعة ويكيبيديا العربية على الشّابكة. وتحتوي هذه المقالات على ٣٧٠ علاقة. تمثلت الخطوة الأولى في تقسيم المقالة إلى جُمل، وفي الخطوة الثانية شُرِحت الجُمل من النّاحية الصرفيّة، وكُشف عن العلاقات الدلالية بطريقة تلقائيّة. ومن خلال مقارنة أزواج الكلمات والعلاقات الدلالية التي تربطها تلقائياً، كان من الممكن حساب مقاييس الاستدعاء والدقة ومقاس اختبار- " لكل علاقة.

ولاحقاً لتقييم القواعد النّحويّة التي بُنِيت، نتجت قيم مقاييس الاستدعاء والدقة ومقاس اختبار- F والتي بلغت ٧٨٪ و ٨٥٪ و ٧٩٪ على التوالي.

الاستنتاجات والتّوصيات

اقترحت هذه الورقة البحثية طريقة لتعريف أو تحديد الأنماط الصّرفية المعجمية التي تساعد في استخراج العلاقات الدلالية بين الأسماء العربية. وقد اعتمد هذا الأسلوب على ذخيرة لغوية مكونة من ٢٠٥٠ مقالة مأخوذة من موسوعة ويكيبيديا العربية. وقد تقدّم شرح الخطوات المختلفة للطريقة المقترحة. وبالنتيجة، نفذت أنماط مُحددة باستخدام منصة نوج (platform NooJ) للتعرف تلقائياً على أزواج من الكلمات والعلاقات الدلالية بين هذه المفردات المحددة. وقد كانت النتائج التي حصلت عليها الدراسة مشجّعة جداً (الاستدعاء: ٧٨٪، والدقة: ٨٥٪، واختبار F- : ٧٩٪) مما يثبت أهمية الأنماط الصرفية المعجمية في الكشف عن العلاقات الدلالية للغة العربية. أما من وجهة نظر مستقبليّة، فأوصت الدراسة بوضع خطط لحلّ الأنماط الغامضة، ومن جهة أخرى اقترحت الدراسة تقديم طريقة للكشف عن الأسماء المركّبة في المستقبل. وفي الختام تأمل الدراسة أن يؤدي تطبيق هذه الأنماط لبناء أنطولوجيا معجميّة للغة العربية.

Natural Language Processing and Information Systems. NLDB 2016. Lecture Notes in Computer Science, vol 9612. Springer, Cham

Year of publication: 2016

Arabic Quranic Search Tool Based On Ontology

Mohammad Alqahtani, Eric Atwell

أداة بحث قرآنية عربية مصمّمة على أساس الأنطولوجيا

المُقدّمة

القرآن الكريم هو نص مقدس باللغة العربية، وهناك حاجة كبيرة لاسترداد المعلومات من القرآن الكريم. ويمكن تصنيف التّقنيات المستخدمة في استرداد المعلومات من القرآن الكريم إلى نَوعين: النوع الأول تقنيات تعتمد على الأساس الدلالي؛ والنوع الثاني هو تقنيات تعتمد على الكلمة المفتاحية. وتُعتبر التقنية المستندة إلى الدلالات أداة بحث تعتمد على المفهوم وتسترد النتائج استناداً إلى معنى الكَلمة أو المطابقة للمفهوم، في حين أنّ التقنية المستندة إلى الكلمات المفتاحية تعمل على استرجاع النّتائج استناداً إلى الأحرف المتطابقة مع مفردات استعلامات الكَلمات. وتستخدم معظم أدوات البحث القرآنية تقنية البحث بواسطة الكَلمات المفتاحية. أما تقنيات البحث الدلالي القرآني المستخدمة حالياً فهي: تقنيات معتمدة على علم توصيف اللغة (الأنطولوجيا)؛ وتقنيات قائمة المرادفات؛ وأخيراً تقنيات استرجاع المعلومات العابرة للغات. تبحث التّقنية القائمة على الأنطولوجيا عن المفهوم أو المفاهيم التي تُطابق استعلام الكَلمات المستخدم، ثم تسترجع هذه التقنية الآيات المتعلقة بهذا المفهوم/ المفاهيم. من جهة أخرى، تُعطي تقنية قائمة المرادفات جميع مرادفات الكلمات المستخدمة في عملية الاستعلام باستخدام نظام(**WordNet**) . بعد ذلك، تعثر على كل الآيات القرآنية المطابقة لمرادفات تلك الكلمات. وتقوم تقنية استرجاع المعلومات العابرة للغات بترجمة كلمات استعلام الإدخال إلى لغة أخرى، ومن ثم تسترجع الآيات التي تحتوي على كلمات تتطابق مع الكلمات المُترجمة.

تستعرض هذه الدراسة وتُصنّف معظم الأنواع الشّائعة لتقنيات البحث التي طُبِّقت على القرآن الكريم، ثم تناقش القيود المفروضة على تلك التّقنيات، كما تستعرض هذه الورقة معظم أنماط التّوصيف القرآنية الموجودة وأوجه القصور فيها. وأخيراً تشرح أداة بحث جديدة تُسمّى أداة البحث الدلالي في القرآن المستندة على أساس علم توصيف اللغة القرآنية. وهذه الأداة تحاول أن تتغلب على جميع القيود المفروضة على تطبيقات البحث القرآنية القائمة حالياً.

المشكلة البحثية

لوحظ وجود العديد من أوجه القصور في عملية استرجاع (retrieval) الآيات القرآنية عند استخدام تقنيات الاستعلام المعتمدة على الكلمات المفتاحية القائمة حالياً. وهذه المشاكل هي بالتحديد: استرجاع آيات غير ذات صّلة بالاستعلام(Query)، أو عدم استرجاع بعض الآيات ذات الصّلة بالاستعلام أو عدم ترتيب الآيات المسترجعة بواسطة الاستعلام. كما تشمل تلك المشاكل أو القيود التي تؤثر على هذه التّقنية المعتمدة على الكلمات المفتاحية: سوء فهم المعنى الدّقيق لكلمات الإدخال المستخدمة في عملية الاستعلام وإهمال بعض نظريات استرجاع المعلومات. ومن هنا، فإن جميع هذه القيود اعتبرت ذات أهمية كبيرة في تصميم أدوات البحث القرآنية العربية الدلالية.

اتبعت هذه الدراسة الأسلوب السّردي، فتناول المبحث الأول الاستعراض المرجعي للدراسات السابقة حول بنية (structure) الآيات القرآنية، وتطبيقات البحث القرآنية والدراسات السابقة المتعلقة بأدوات البحث القرآني وأساليب أو أنماط التوصيف القرآنية. في حين ناقشت الدراسة في المبحث الثاني منهجية البحث القرآني المعتمدة على الأنطولوجيا، وأخيراً ختمت بالاستنتاج والتوصيات.

أسلوب البحث

أداة البحث الدلالي القرآني العربية (Arabic Quranic semantic search tool)

يمثّل هذا المبحث المحتوى الرّئيسي من هذه الدراسة، ويعرض إطاراً لأداة بحث دلالي جديدة أُطلق عليها اسم «أداة البحث الدلالي القرآني العربية المصمّمة على أساس الأنطولوجيا». وتهدف أداة البحث هذه إلى توظيف كل من تقنيات استرجاع (retrieved) المعلومات وتقنيات البحث الدلالي. وقد صممت هذه الأداة اعتماداً على عدد من الدراسات السابقة ذات الصّلة. وتنقسم «أداة البحث الدّلالي القرآني العربية المصمّمة على أساس الأنطولوجيا» إلى ستة مكونات: (١) توصيف لغة القرآن الكريم (Quranic Ontology)، (٢) قاعدة البيانات القرآنية(Quranic Database)، (٣) محلل اللغة الطبيعية (Natural Language Analyser)، (٤) نموذج البحث الدلالي (Semantic Search Model)، (٥) نموذج البحث عن الكلمات المفتاحيّة(Keyword Search Model)، (٦) نموذج التصنيف والترتيب (Scoring and Ranking Model). ويحتوي (علم توصيف اللغة القرآنية) على تصنيف قرآني محاذ جديد. أمّا (قاعدة البيانات القرآنية) فتتكوّن من: نص القرآن الكريم باللغة العربية؛ وثماني ترجمات إنجليزية للقرآن الكريم؛ وأربعة تفاسير مختلفة للقرآن الكريم؛ ومعجم معاني مفردات القرآن؛ وأسباب النزول، ومفاهيم قرآنية، والكيانات (الهيئات) المسمّاة (Named Entities) على أساس النّطاق القرآني. ويخضع طلب المستخدم للاستعلام لعدة عمليات مختلفة حسب نموذج (محلّل اللُغة الطبيعية) المقترح، حيث يقوم المحلل بتحليل طلب استعلام اللغة الطبيعية ثم يطبق تقنيات البرمجة اللغوية العصبية المختلفة على طلب الاستعلام المرمّز. وتشمل هذه التّقنيات: التّصحيح الإملائي (spell correction)، وإزالة كلمات التوقف (stop word removal)، واقتفاء جذر أو أصل أقسام الكلام temming and Part Of Speech (POS) tagging)). بعد ذلك، يستخدم محلل اللغة الطبيعية برنامج (Arabic WordNet) لتوليد مرادفات لكلمات الاستعلام التي تمت إعادة صياغتها، ثم يضيف المحلل علامات دلالية لهذه الكلمات باستخدام قائمة الكيانات أو الهيئات المسماة، ومن ثمّ يقوم بإرسال النتائج إلى نموذج البحث الدلالي من جهة أخرى، يقوم نموذج البحث الدلالي بالبحث في قاعدة البيانات القرآنية بواسطة (SPARQL) للعثور على المفاهيم المتعلقة بالاستعلام التّطبيعي (normalised query) ومن ثم استرجاع النتيجة إلى نموذج التّصنيف والترتيب. وإذا لم يتم العثور على نتيجة يقوم نموذج البحث بواسطة

الكلمات المفتاحية بالبحث عن الآيات القُرآنية التي تحتوي على كلمات مطابقة لكلمات الإدخال التي تمّ تحليلها.

ويقوم نموذج التّصنيف والترتيب بتصفية النتائج المسترجعة من نموذج البحث الدلالي ونموذج البحث بواسطة الكلمات المفتاحية؛ من خلال استبعاد الآيات المكررة. ثم يقوم نموذج التّصنيف والترتيب بعد ذلك بترتيب وتصنيف النتائج المسترجعة على أساس عدد الكلمات المطابقة في كل من: النتائج والكيانات (الهيئات المسمّاة) لكل من السؤال والجواب والمسافات القصيرة بين التعبيرات المتطابقة في النتائج المسترجعة وكلمات السؤال. وأخيراً، يقدّم نموذج التصنيف والترتيب النتائج للمستخدم، ثم يسجل النتيجة التي تمّ اختيارها.

الاستنتاجات

تلخّص هذه الدراسة تقنيات البحث المستخدمة حالياً في أدوات البحث القرآنية، فقد ناقشت هذه الورقة البحوث والدراسات السابقة التي أجريت على أساليب البحث القرآنية وتوظيف أطولوجيا اللُغة فيها. ووفقاً لهذه الدراسة، هناك العديد من التحديات والمشاكل التي تواجه هذه التقنيات ومن ضمنها، أولاً: القيود المفروضة على أدوات البحث القرآني المستخدمة حالياً لاسترداد جميع المعلومات المطلوبة، فهذه الأدوات لا تقدم للمستخدمين البحث عن طريق المفاهيم والعبارات والجمل والأسئلة أو الموضوعات. ومعظم أدوات البحث لا تعمل على تحليل نصوص الاستعلام من خلال تطبيق تقنيات البرمجة اللغوية العصبية، مثل التحليل اللغوي والتدقيق الإملائي. ثانياً: عدم توفر موارد دقيقة وشاملة لأدوات توصيف لغوية إسلامية. فقوائم البيانات القرآنية الموجودة تتخذ نطاقات وأشكالا مختلفة، ولا تتبع معايير توصيف لغوية معيّنة. كما أنّ بعض أنماط التوصيف اللغوية القرآنية ليست متاحة للاستخدام.

ويتركز البحث عن الكيانات (الهيئات) المُسمّاة في اللغة العربية على «اللغة العربية الحديثة». كما أنه لا توجد قوائم للكيانات (الهيئات) مُنسّقة بشكل جيد ومتخصصة بالنص القرآني، مثل أسماء الله الحسنى وأسماء الأنبياء وأسماء الحيوانات، والأوقات أو الأزمنة أو الدين، إلخ.

فاللغة العربية هي لغة مُصرّفة (inflected language) وتمتاز بالإملاء المُعقّد(complicated orthography). وأخيراً، تُعدّ جميع هذه القيود ذات أهمية كبيرة في تصميم أدوات البحث القرآنية العربية الدلالية.

LREC 2014 Proceedings. Ninth International Conference on Language Resources and Evaluation (LREC'14), 26-31st May 2014, Reykjavik, Iceland

Year of publication: 2014

Computational ontologies for semantic tagging of the Quran: A survey of past approaches

Sameer M. Alrehaili, Eric Atwell

الأنطولوجيا الحاسوبية للوسم الدلالي في القران الكريم: دراسة استقصائية للمنهجيات السابقة

لم تقتصر التطورات في مجال معالجة اللغات الطبيعية (Natural Language Processing) واستخراج النصوص (Text Mining) على الجوانب العلمية أو اللغوية في النصوص العربية، وإنما امتدت لتشمل النصوص الدينية بما فيها القران الكريم والأحاديث النبوية الشريفة، كونهما أهم مصادر التشريع الإسلامي. فكان لعلم الأنطولوجيا (ontology) الحظ الأوفر في المساهمة في تحسين المعرفة في مثل هذه النصوص، من حيث المجالات الدلالية. لقد ورد تعريف الأنطولوجيا في الذكاء الاصطناعي(Artificial Intelligence) بأنه علم ذو مواصفات معينة تساعد البرامج والبشر على عملية تبادل المعرفة بطريقة تفاعلية، حيث يقوم هذا العلم على معاني الكلمات والنصوص ودلالاتها.

ومن الملاحظ أن هناك اختلافات جوهرية بين العلماء والفقهاء في دلالات بعض الكلمات أو الآيات في القران الكريم والأحاديث النبوية الشريفة، مما أدى إلى اختلاف التحليلات الدلالية للنصوص الدينية و للقران الكريم تحديدا، فكان من الضروري تصميم شكل موحّد لهذه التحليلات لتسهيل العملية على المستخدم وإعطاء دلالات واضحة وموحّدة، لكن ذلك ليس بالأمر السهل. فجاءت هذه الدراسة محاولة لإيجاد أفضل وسم قد جرى سابقا عن طريق جمع مجموعة من الوسم والتحليلات الدلالية ومقارنتها مع بعضها للخروج بأفضل تحليل

وتوفيرها مجاناً على شبكة الإنترنت، فاعتمد الباحثون على مجموعة من المعايير كمقياس لتحديد الجودة، أهمها: لغة نص القران الكريم الذي تمت دراسته (سواء أكان من النصُّ الأصليُّ في اللغة العربية أم نصا مترجما من لغات أخرى)، والمواضيع التي غطّتها الدراسة (مثلا موضوع أسماء الحيوانات فقط)، ونسبة تغطية الدراسة (القران الكريم كاملا، جزء منه أم موضوع معين فقط)، والتكنولوجيا التي استُخدمت في عملية التحليل، وعدد المفاهيم المستخدمة، ومدى توفر هذه البيانات الموسومة بشكل مجاني أو مدفوع، كذلك نوع العلاقات بين المفاهيم سواء كانت جزءا من كل أو مترادفات أو متناقضات أو تابعة، وأخيرا الطريقة المعتمدة في عملية التحقق وهي التي تحكم على جودة العمل، وذلك عن طريق خبراء في هذا المجال أو من مصادر علميّة موثوقة.

وبعد القيام بمجموعة من الدراسات والأبحاث تبين أن معظم عمليات الوسم كانت غير مكتملة أو مركّزة على مجال محدد جدا، فليس هناك إجماع واضح على بعض الدلالات في القران الكريم أو في الطريقة التي تم استخدامها في بناء هذه البيانات الموسومة و التحقق منها.

11th International Computer Engineering Conference (ICENCO)

Year of publication: 2015

Ontology-based model for Arabic lexicons: An application of the Place Nouns in the Holy Quran

Waseem Alromima; Ibrahim F. Moawad; Rania Elgohary; Mostafa Aref

نموذج لمعاجم اللغة العربية معتمد على الأنطولوجيا: تطبيق على أسماء الأماكن في القران الكريم

أصبح الأنترنت المصدر الأساسي للمعلومات التي نحصل عليها يوميا، ويساعد علم الأنطولوجيا بتبادل هذه المعلومات بين المستخدمين. يعدُّ علم الأنطولوجيا القائم على النموذج الدلالي أحد العلوم الأساسية التي تقوم عليها العديد من المجالات، ومن أهمها: استرجاع المعلومات (**Information Retrieval**) وعلم الويب الدلالي((**semantic web**إضافة إلى إدارة المعرفة **Management** (**Knowledge**) . كما أن هناك أهمية لاستخدام الأنطولوجيا في العلوم الدينية والبحوث اللغوية وتطبيقات الويب الدلالية (**Semantic Web applications**). لقد كانت العلوم الدينية واحدة من العلوم التي وظّفت الأنطولوجيا وساعدت في تطوير المعرفة في النصوص الدينية خاصة القرآن الكريم والحديث الشريف. وهناك أسباب عدة تكمن في اختيار مفردات القرآن الكريم بشكل خاص، ومن أهمها أن لغة القرآن الكريم تعتبر أوضح أشكال اللغة العربية وأنقاها وأكثرها شهرة.

تعرض هذه الورقة البحثيّة طريقة بناء الأنطولوجيا التي تمر عبر أربع مراحل مختلفة: المرحلة الأولى متمثلة باستخراج الأنطولوجيا (**Ontology Extraction Phase**)، وتهتم باستخراج المفاهيم المتعلقة بأسماء المكان من موارد المعرفة المختلفة مثل الدراسة اللغوية الحاسوبية(**computational study linguistic**) من مشروع القرآن الذي وضعته جامعة ليدز. أما المرحلة الثانية فهي مرحلة التصميم والتكامل (**Design and integration Phase**)

وتتمثل هذه المرحلة بتصميم هيكل الأنطولوجيا المطلوبة الذي يعتمد بشكل أساسي على المفاهيم التي استُخرجت في الخطوة الأولى. وتهدف هذه الخطوة إلى بناء الأنطولوجيا بشكل متكامل من الأسماء التي ذُكِرت في القرآن الكريم. لقد استخدم الباحثون نموذجا يُسمّى لغة النمذجة الموحدة (UML Unified Modeling Language) يقوم بدمج المصطلحات المتشابهة وتصنيفها. فمثلاً مفهوم (الموقع الجغرافي) يندرج تحته أربع فئات، هي: البلد، المدينة، الجبل، الوادي. أما المرحلة الثالثة التي تسمى بمرحلة التحقق(Verification Phase)، فتهتم بالتحقق من المفاهيم، ويقوم بهذه المهمة خبراء مختصون في المجال، حيث يقومون بمراجعة المفاهيم جميعها الواردة في النموذج وحذف المكرر منها، وكذلك التأكد من تصنيفات المفاهيم في الفئات الصحيحة. وآخر مرحلة هي مرحلة التنفيذ Phase (Implementation)، وهي المسؤولة عن تنفيذ تلك الأنطولوجيا باستخدام اللغة القياسية للشبكة الدلالية، وتعبر عن المفاهيم(concepts) التي سبق تمثيلها في مرحلة التصميم كمدخلات لهذه المرحلة. كما اقتُرِحت أنطولوجيا مختلفة لتمثيل المفاهيم وتصنيفها في القرآن الكريم.

وقد أظهرت النتائج الأولية في هذه الدراسة أن استخدام الأنطولوجيا قادر على تمثيل الكلمات الدلالية التي تسهّل عملية التحليل الدلالي لكلمات اللغة العربية عامة ومفردات القران الكريم خاصة.

Proceedings ICWIT 2012, pp. 193–202.

Year of publication: 2012

Automatic construction of ontology from Arabic texts

Ahmed Cherif Mazari, Hassina Aliane, and Zaia Alimazighi

البناء الآلي للأنطولوجيا باستخدام النصوص العربية

تقوم فكرة هذا البحث على بناء أنطولوجيا لتوصيف اللغة العربية، حيث قدم الباحثون نهجاً للبناء التلقائي للأنطولوجيا يستخدم التقنيات الإحصائية لاستخلاص عناصر الأنطولوجيا مثل المفاهيم والعلاقات من النصوص العربية. وللقيام بذلك استخدموا تقنيتين؛ الأولى هي «التقسيم المتكرر" (repeated segments) لتحديد المصطلحات ذات الصلة التي تدل على المفاهيم المرتبطة بالمجال، والثانية هي تكرار الحدوث (co-occurrence) وتُستخدم للربط بين هذه المفاهيم المستخلصة من علم الأنطولوجيا عن طريق العلاقات الهرمية أو غير الهرمية. وقد اختُبِر النموذج المقترح اعتمادا على مجموعة من النصوص العربية التي أعِدّت مسبقاً، ثم معالجتها لإزالة الغموض وبعض الزوائد. وتكمن أهمية الأنطولوجيا المقترحة في أنّها أداة داعمة في معالجة اللغات الطبيعية التي تحلل النصوص، كما يمكن استخدامها لاستخراج المعلومات من الذخائر اللغوية العربية.

قُسّمت منهجية البحث إلى ثلاث خطوات رئيسة: الخطوة الأولى تناولت عملية تحضير الذخيرة اللغوية (preparation of the corpus)، حيث إن بناء الأنطولوجيا يتطلب جمع معلومات لتكون مورداً خلال عملية بناء النموذج وتطويره، وبالتالي، فإن البيانات التي تُجمع تتناول نوع الذخيرة (Corpus type)، وحجمها (Corpus size) والمواضيع المرتبطة بها (Topics) فالبيانات التي تُجمع تسهم بشكل أساسي في نجاح المنهجية المقترحة. وبعد جمعها قام الباحثون بمعالجتها عن طريق حذف الحروف غير العربية والأرقام وكلمات الربط (Stop

Words)، كذلك إزالة الحروف الخاصة والأرقام، بالإضافة لإزالة الزوائد السابقة واللاحقة عن طريق تجذيع الكلام(Light stemming).

أما الخطوة الثانية فقد تضمّنت الاستخلاص التلقائي للمفردات المرشحةAutomatic extraction) of candidate terms). في هذه الخطوة تُستخلص المفردات التي ستشكل الأفكار الموجودة في الأنطولوجيا باستخدام طريقة التقسيم المتكرر اعتمادا على مبدأ أن المفردات المهمة والمؤثرة يجب أن تتكرر عدداً من المرات داخل النص، وبذلك نحصل على أزواج من المصطلحات التي ترد بصورة أكثر تكرارا (co-occurrence) في المجموعة. وتوفر لنا نتيجة هذه المعالجة قائمة بأزواج المصطلحات التي ستستخدم لتحديث علم الأنطولوجيا. ولذلك، فإن الهدف من هذه الخطوة هو تحديد المصطلحات التي تدل على المفاهيم المتعلقة بالفكرة، ثم التحديد من بين هذه المصطلحات الأزواج التي لها صلات مع عناصر من الأنطولوجيا المستخلصة.

تضمنت الخطوة الأخيرة عملية تحديث الأنطولوجيا الذي تم توليدها، اعتماداً على مقارنة المصطلحين المرشحين المستخرجين بالتسميات الواردة في مفاهيم الأنطولوجيا. وفي النهاية قام الباحثون باختبار النهج المقترح باستخدام لغة البرمجة بايثون (Python)، نظرا لكفاءتها في علوم معالجة اللغات الطبيعية.

International Journal on Islamic Applications in Computer Science and Technology (2016)

Year of publication: 2016

Towards Concept Extraction for Ontologies on Arabic language

Abeer Al-Arfaj and Abdul Malik Al-Salman

نحو استخراج المفاهيم للأنطولوجيا في اللغة العربية هنايا غسان

تعدّ الأنطولوجيا (Ontology) أو علم التوصيف واحداً من أهم نماذج التمثيل المستخدمة لتمثيل المعرفة وإعادة استخدامها. وفيما يتعلق باللغة العربية، تتميز هذه اللغة بالتعقيد الصرفي والقواعدي والدلالي وبأنها لغة اشتقاقية للغاية، مما يجعل التحليل الصرفي مهمة معقدة. لهذا، فإن استخراج المصطلحات العربية تلقائياً ليس سهلاً. بالإضافة إلى ذلك، فإن استخراج المفاهيم من الوثائق العربية أكثر صعوبة. وبالتالي، فإن أدوات معالجة اللغة الطبيعية التي تم تصميمها للغة الإنجليزية ليس بإمكانها تلبية احتياجات اللغة العربية، كما أن اللغة العربية تفتقر إلى خاصيّة ابتداء الكلمات بأحرف كبيرة (Capitalization) مما يزيد في تعقيد عملية استخراج الأسماء العربية.

في هذه الورقة البحثية، تعامل الباحثان مع أساسيات بناء توصيفات النصوص، كما قاما بدراسة مشاكل استخراج المفاهيم من نصوص النطاق (domain text) العربية. من ناحية أخرى، قام الباحثان بعمل مقارنة بين أهم التقنيات الحالية المستخدمة لاستخراج المصطلحات العربية مع تسليط الضوء على التحديات التي تفرضها خصائص اللغة العربية. وساهمت هذه الورقة البحثية في توفير تحليل شامل للطرائق الآلية لاستخراج المصطلحات وعرض بعض التعريفات الأساسية المرتبطة بالمصطلحات، وعمل تلخيص خاص بطرائق استخراج المصطلحات العربية الحالية مع تسليط الضوء على نقاط القوة والضعف لكل طريقة، واقتراح توجيه بحث مستقبلي جديد لاستخراج المصطلح العربي.

وأجريت مقارنة بين طرق استخراج المصطلحات العربية الحالية التي صنفت بناءً على نوع الاستخراج وتقنية الترشيح. أثبت تحليل الباحثين أن الغالبية العظمى من الدراسات ذات الصلة قامت بتطبيق تحليل لغوي ضحل، وأن بالإمكان تحسين النتائج باستخدام تطبيقات لغوية أوسع. وخلص الباحثان من هذه المقارنة إلى أن عملية استخراج المصطلحات العربية هي مهمة معقدة، وأن اختيار النهج المناسب يعتمد على نوع موارد البيانات المتاحة وعلى التطبيق.

بالرغم من الجهود التي بُذلت بهدف دمج التدابير الإحصائية لاستخراج المصطلحات العربية، إلا أن الغالبية العظمى من تقنيات استخراج المصطلحات العربية غير قادرة على إنتاج مصطلحات نادرة (rare terms) . وثمة قيد آخر مع الطريقة الحالية لاستخراج المصطلحات العربية هو أن تلك الطريقة تقوم باستخراج المصطلحات العامة التي لا صلة لها بنطاق (domain) معين. لذلك ينبغي استخدام موارد المعرفة الخاصة بالنطاق لدعم أساليب استخراج المصطلحات. وبسبب محدودية المعرفة الخاصة بالنطاق العربي، تواجه طرق استخراج المصطلحات العربية تحديات في استخراج المصطلح العربي من نصوص خاصة بالنطاق.

خلص الباحثان إلى أنه استناداً إلى الأدبيات، ظهر أن النهج اللغوي والإحصائي يعانيان من بعض نقاط الضعف عند استخدامهما بمفردهما. والنهج الإحصائي غير قادر على تحديد المصطلحات النادرة؛ فهو يركز على الخصائص الإحصائية للمصطلحات ويتجاهل المعرفة اللغوية والدلالية. ومن ناحية أخرى، النهج اللغوي يعتمد على اللغة ولا يمكن أن يقوم بالقياسات الصحيحة في مجموعات البيانات الكبيرة. ولتفادي نقاط الضعف تلك والاستفادة من مزايا كل طريقة، فقد طبقت معظم الأبحاث العربية طريقة هجينة لاستخراج المصطلحات العربية.

11th International Business Information Management Association Conference (2009)

Year of publication: 2009

Using an Arabic ontology to improve the Q/A task

Lahsen Abouenour, Karim Bouzoubaa

استخدام الأنطولوجيا في اللغة العربية لتحسين مهمة السؤال والجواب

يحتاج المستخدمون في محركات البحث إلى أدوات متقدّمة تسّهل عليهم الوصول إلى المعلومات، ومن طرق البحث عن المعلومات واسترجاعها نظام السؤال والإجابة (Question Answering) الذي يمتاز بتوفير الجهد، فيقوم المستخدمون بالبحث عن إجابة دقيقة عن سؤال بدلاً من مراجعة مجموعة من المستندات والبحث فيها، ويمكن تحسين ذلك عن طريق توسيع الاستعلام الدلالي (semantic Query Expansion) كما في هذا البحث. تتألف دورة نظام السؤال والإجابة من ثلاث مراحل رئيسة، تبدأ بالمعالجة المسبقة للسؤال المدخل، ثم استرجاع الوثائق المرشّحة التي تتضمن إجابة هذا السؤال، وانتهاءً بمعالجة كل وثيقة من الوثائق المرشّحة بنفس الطريقة التي تتم بها معالجة السؤال لاسترجاع تلك الجمل التي قد تحتوي على الإجابة.

لقد حاول الباحثان في هذه الورقة تحسين إحدى مراحل نظام الأسئلة والإجابة، وهي مرحلة استرجاع الوثائق المرشحة في عملية البحث باللغة العربية، واعتمد نهجهما على توسيع قائمة الكلمات المفتاحية للسؤال المدخل ليس فقط باستخدام الاشتقاقات الصرفية للكلمات المفتاحية للسؤال، ولكن باستخدام الدلالة أيضا. وللاستعلام عن سؤال ما، استخدم الباحثان النموذج المسمّى(Amine Arabic WordNet Ontology (AAWN)) الذي يهتم بالتسلسل الهرمي للمفاهيم مع مرادفاتها وتعريف كل منها.

يقوم نموذج AAWN الذي استخدمه الباحثان بالتوسيع الدلالي للاستعلام على مبادئ أربعة: المبدأ الأول يقوم على معالجة مرادفات (Synonyms) الكلمة المفتاحية، فعلى سبيل

المثال كلمة (منظّمة) ترادفها كلمة (تنظيم)، والمبدأ الثاني يقوم على أن الكلمة المفتاحية لها نوع عام (super type) حسب التقسيم الهيكلي لتلك الكلمة، فمثلا كلمة (منظّمة) تنتمي للكلمة المفتاحية الأعلى مستوى (جماعة) . أما المبدأ الثالث من مبادئ AAWN، فيعتمد على تعريف (Definition) الكلمة المفتاحية، والمبدأ الأخير يقوم على توسيع الاستعلام عن طريق الكلمات الفرعيّة (Subtype) المنبثقة من الكلمة المفتاحية، على سبيل المثال كلمة (منظّمة) ترتبط بالكلمات (اتحاد، مؤسسة، جمعية ....)

ولإثبات فعالية النموذج المقترح قام الباحثان باستخدام تجربتين: الأولى يدوية والثانية آلية. في التجربة الأولى، أخذ الباحثان مجموعة من الأسئلة واستعلما عن إجاباتها مرة من خلال استخدام نموذج (AAWN) والمرة الثانية دون استخدامه، فلاحظا تحسناً واضحاً في نتائج البحث عند استخدامه. أما في التجربة الثانية فكانت العملية آلية باستخدام محرك البحث ياهو مع نظام جافا لاسترجاع المعلومات(Java Information Retrieval System (JIRS) ، وقد أثبتت تلك التجربة تحسناً في النتائج عند استخدام النموذج المقترح. هنا

Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta

Year of publication: 2010

Al-Khalil: The Arabic Linguistic Ontology Project

Hassina Aliane, Zaia Alimazighi, Mazari Ahmed Cherif

الخليل: مشروع توصيف (أنطولوجيا) اللغويات العربية

لقد تزايدت في الآونة الأخيرة الأعمال التي تركز على توفير الموارد اللغوية المحوسبة والأدوات والتطبيقات المخصصة للتعامل مع اللغة العربية. والمشروع الذي تقدمه هذه الورقة يهدف إلى بناء توصيف أو ما يسمى بالأنطولوجيا (ontology) للغويات العربية، فالأنطولوجيا هي علم يهتم بالتوصيف الرسمي لمجال معين (domain) من خلال تحديد المفاهيم الخاصة به (concepts) وصفاتها (attributes) والعلاقات بين هذه المفاهيم (relations)، وعادة ما تكون هذه المفاهيم منظمة بشكل هرمي (hierarchy). أما المنهجية المطّبقة في هذا البحث فتتمثل بمناقشة تطوير الأنطولوجيا وتقديم الاقتراحات حول استخدامها لتصميم الأدوات والموارد والبرمجيات المفيدة لكل من اللغويين والباحثين في مجالات معالجة اللغات الطبيعية واسترجاع المعلومات. وقد سُمّي النموذج المقترح بـ «الخليل" (Al –Khalil) ، وركّز على مفاهيم قواعد اللغة العربية التقليدية التي لا تظهر في النظريات اللغوية الأخرى.

يعود سبب تسمية هذا المشروع إلى النحوي الشهير الخليل بن أحمد الفراهيدي، لأنه يُعدُّ من أوائل الذين اهتموا ببناء أنطولوجيا خاصة باللغة العربية في كتابه «العين». ولقد طُوِّر هذا النموذج بالاعتماد على أنطولوجيا لغوية موجودة أصلا وهي أنطولوجيا غولد (General Ontology for Linguistic Description) واختصارا تكتب "GOLD"، وهي أول أنطولوجيا صُممت للوصف اللغوي المبني على وصف الشبكة الدلالي (The Semantic Web)، وهكذا فإن مشروع الخليل طُوِّر بخطوتين رئيسيتين هما:

توسيع هذه الأنطولوجيا يدويا عن طريق إضافة المفاهيم اللغوية العربية وربطها بمفاهيم غولد.

استخدام خوارزمية الاستخراج التلقائي لاستخراج مفاهيم جديدة من النصوص اللغوية العربية لإثراء الأنطولوجيا، وتعتمد هذه الخوارزمية على الحسابات الإحصائية لتكرار الكلمات وجذورها.

وقد اعتمد النموذج المقترح على مجموعة من المعاجم العربية لاستخراج المفاهيم الأكثر وضوحا من قواعد اللغة العربية التقليدية ومن ثم ربط هذه المفاهيم بمفاهيم غولد، ثم تطبيق خوارزمية الاستخراج التلقائي عن طريق إجراء العديد من الطرق الإحصائية على الأجزاء المتكررة جنباً إلى جنب مع بعض خطوات معالجة النصوص التي تشمل: تقطيع النص (segmentation)، والتجذيع الخفيف (light stemming)، وإزالة كلمات التوقف (stop words removal). ولتطبيق هذه المهمة استُخدمت لغة برمجة بايثون (Python). وفي نهاية المطاف عُرِضت المفاهيم والعلاقات المرشّحة على خبراء في اللغة قبل إدراجها في الأنطولوجيا، بالإضافة إلى توفير نسخة للتصفح وتحرير الواجهات للاختبار والتحقق من صحة النسخة الأولى من الأنطولوجيا المقترحة قبل وضعها على الإنترنت.

Proceedings of the 2015 International Conference on Soft Computing and Software Engineering (SCSE'15), Procedia Computer Science, Volume 62, 2015, Pages 513-520

Year of publication: 2015

Arabic Ontology Model for Financial Accounting

A.Hegazy, M.Sakre, Eman Khater

نموذج الأنطولوجيا العربية للمحاسبة المالية

يعد بناء نظام ذكي للمحاسبة من الأمور المهمة لإيجاد نموذج مفاهيمي (**conceptual model**) يشتمل على جميع المفاهيم والاصطلاحات ويبين العلاقات بينها من خلال ابتكار الأنطولوجيا. هذا النوع من النماذج يُستخدم للتمييز بين عدة أنواع من الإيرادات والنفقات وتمثيلها في البيانات الماليّة. وعلاوة على ذلك، يمكن استخدامه في تحديد المفاهيم ذات الصلة وفهمها والطريقة التي تستخدم هذه المفاهيم في العمليات المالية. وقد زادت أهمية أدوات معالجة اللغة العربية بشكل كبير في العقد الماضي بسبب الزيادة الهائلة في المحتوى الرقمي العربي على شبكة الإنترنت، وفي عدد مستخدمي الإنترنت الذين يتحدثون اللغة العربية. هذه الحقيقة تزيد من أهمية إنشاء أدوات معالجة اللغة التي يمكنها معالجة هذا المحتوى، والتفاعل مع هؤلاء المستخدمين بأفضل الطرق. يقوم علم الأنطولوجيا (في تكنولوجيا المعلومات) على تقسيم المتغيرات اللازمة لمجموعة من الحسابات ويحدد العلاقات بينها، فهي بحد ذاتها عملية تصنيف لهذه المتغيرات.

تعاني اللغة العربية من محدودية توفر مثل هذا النوع من الأنظمة، فأتمتة اللغة العربية وتطويرها يحتاج أكثر مما تقدمه الكلمات الرئيسية أو قواعد اللغة العربية التقليدية في مجال مفاهيم المحاسبة. ومع ذلك، فإن محرّكات البحث المتاحة التي تدعم اللغة العربية تقتصر عادة على عمليات البحث عن الكلمات الرئيسية ولا تأخذ بعين الاعتبار الدلالات الأساسية

للمحتوى. في هذا البحث، اقترح الباحثون نموذجا لتمثيل المعرفة المحاسبية المالية العربية في مجال تكنولوجيا الحاسوب باستخدام نماذج الأنطولوجيا. تتناول هذه الورقة البحثية علم الأصول المالية العربية على وجه الخصوص، وقد هدف الباحثون أساسا في عملهم هذا إلى بناء علم الأنظمة المحاسبيّة كخطوة أولى لإنشاء مستودع محاسبيّ تنظيميّ يسمح بتخزين المعرفة ونشرها في هذا المجال. كما قدّمت هذه الورقة البحثية نتائج تحليل نظم المعلومات المحاسبيّة الماليّة من وجهة نظر علم الأنطولوجيا، وركز التحليل على النظم والتمييز بين الأصول والخصوم والإيرادات والمصروفات في بعض الحالات.

وفي هذه الورقة البحثية أُعِدَّت بعض المفردات بواسطة خبراء في مجال المحاسبة وخاصة المحاسبة المالية. وكذلك استُكشفت العلاقة بين المفاهيم والتحقق منها، فاعتمد الباحثون على استخدام خوارزمية التصنيف المركزي (Centroid Clustering) لتبسيط النظام المحاسبي وفهمه بطريقة أفضل وأكثر واقعية. وتبين النتائج بعد القيام ببعض الاختبارات أن تقنية التسلسل الهرمي أو التصنيف المركزي للمفهوم هي تقنية موثوقة لإقامة علاقة قوية بين المفاهيم كافة وربطها ببعضها.

International Symposium on Web Services, Zayed University. Dubai: April (pp. 9-10).

Year of publication: 2008

Ontology-Based Semantic ConteFramework (OBSC) Framework for Arabic Web Contents

.AlKhatib,B., Kawas, M., Bshara، W., TalalKalla, M

إطار السياق الدلالي القائم على علم الأنطولوجيا لمحتويات الويب العربية

يمكن الوصول إلى المحتويات العربية بجميع أشكالها على شبكة الإنترنت سواءً كانت على شكل وثائق نصية (word documents) أو بلغة إعداد النصوص (HTML) أو ملفات نوع (PDF) وغيرها. من هنا فإن مهمة إطار السياق الدلالي القائم على علم الأنطولوجيا (OBSC) هي تحويل أيٍّ من هذه الأشكال إلى هياكل مفاهيمية، بحيث يمكن لجهاز الحاسوب فهمها واستخدامها في التطبيقات الدلاليّة. مثل محرّكات البحث الدلاليّ والموسوعة الدلاليّة والقواميس الدلاليّة وغيرها. يجب أن يكون هذا الإطار قادراً على التعامل مع محتويات الويب العربية باستخدام معجم المفردات العربية، ولكن معظم الأبحاث السابقة كانت تهمل المعاني الدلاليّة وتهتم بنظم استرجاع المعلومات والنهج الإحصائيّ.

يعتمد نظام (OBSC) على معجم المفردات العربية (Arabic WordNet) فكل كلمة يمكن أن تدل على اسم، أو فعل، أو صفة، أو ظـرف، وهـذا مـا يسمى بأقسام الكلام (part of speech) ، فمثلا كلمة (درس) يمكن أن تدل على فعل أو اسم باختلاف الحركات لكن في كلا الحالتين لها الحروف نفسها. لقد صُمم علم الأنطولوجيا لربط هذه الكلمات بالعلاقات الدلاليّة، مثل الأصل(hypernym) والفرع (hyponym) فمثلا النسر، الصقر، البلبل هذه فروع، وتقع جميعها تحت مسمى واحد (طيور) وهو (الأصل) و الكناية (metonym) أو غيرها من الروابط.

يقوم إطار OBSC على مجموعة من الخطوات التي من شأنها تطوير بعض التطبيقات المتعلقة بالدلالات. الخطوة الأولى هي الترميز بتقسيم النص إلى كلمات (Tokenization)، وفيها تُأخذ مجموعة من الكلمات من أيّ نص عربي، ثم تُكتب هذه الكلمات واشتقاقاتها، ومن ثم تُجمع في قائمة وهو ما يسمى بالفهرسة (Indexing) فعندها يمكن استرجاع أي كلمة أو إحدى اشتقاقاتها بغض النظر عن نوعها من أقسام الكلام. أما الخطوة الثالثة فتتمثل بإزالة الغموض عن الكلمات (Word Sense Disambiguation) التي لها الكتابة نفسها وتنتمي إلى النوع نفسه، لكن بمعنى مختلف، فمثلا، كلمة « العالَم» بالفتحة اسم، و « العالِم» بالكسرة هي أيضا اسم، ففي هذه الحالة نحتاج إلى إزالة الغموض عن طريق اتباع الحدس لإيجاد المعنى الأقرب للكلمة في سياق معيّن. أما الخطوة الأخيرة فهي التجميع Clustering)) وتُستدخل فـي هذه الخطوة جميع المحتويات المفاهيميّة، عندها يمكن للمستخدم الحصول على وثائق مشابهة ومنظّمة بشكل سريع وسهل.

ولاختبار هذا الإطار طُوّرت موسوعة تحت اسم (Arapedia) حيث يمكن إضافة أيّة معلومة لهذه الموسوعة، ويمكنها ترجمة أيّة كلمة غامضة (تحتمل أكثر من معنى) بناءً على البحث الدلاليّ للكلمة في السياق، فمثلا إذا بحثنا في عبارة « معدن الذهب» فإنه مباشرة يترجم كلمة ذهب بمعنى «gold" وعند كتابة العبارة " ذهب الرجل" يترجمها "went" بناءً على السياق، فأثبت هذا الإطار جدارة وكفاءة عالية.

٣-٣-٥ أبحاث شبكات الكلمات

تضم ملخصات ستة أبحاث بينها ملخصان من النوع (أ) يعطيان السمات العامة لمشروع شبكة الكَلمة العربيّة «ووردنيت» بشكل عام والوضع الرّاهن والتوسعات المستقبليّة ، بينما تتضمن أربعة ملخصات أبحاث من نوع ب تشمل التحديات التي يواجهها المشروع من جهة اللغة العربيّة وإثراء العلاقات الدلالية في «ووردنيت» العربية من خلال أنماط معجمية وصرفية ومحاولة استخلاص أوتوماتيكي للمرادفات «ووردنيت» القرآنية واستخدام ووردنيت العربية للفهرسة الدلالية في نظام استعادة المعلومات.

Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), May 2006, Genoa, Italy

Year of publication: 2006

Building a WordNet for Arabic

Sabri Elkateb, William Black, Piek Vossen, Horacio Rodríguez, Adam Pease, Musa Alkhalifa, Christiane Fellbaum

بناء شَبكة الكَلمة العَربيّة "وردنيت "

المقدّمة

تستعرض هذه الورقة البحثيّة مشروعاً أُطلق مُؤخراً، يُركّز على بناء مورد معجمي للُغة العربية القياسيّة الحديثة، وذلك اعتماداً على شكة الكلمات الإنجليزية برينستون (**Princeton WordNet**) واسعة الانتشار. وغرضُنا هو تطوير مورد لغوي ذي أساسٍ دلالي اصطلاحي عميق، من أجل إثراء اللُغة العربيّة. ولقد تمّ إنشاء الشّبكة العربية (**Arabic WordNet**) وفقاً لنفس الأساليب المتّبعة في تطوير شبكة (**Euro WordNet**). وبالإضافة إلى عرضها للمعاني والدلالات، تضمّنت الشّبكة تعريف معاني الكلمات بواسطة دلالات مفهومة آلياً في النظام الأول لعلم أصول الكلام، حيث كان الأساس الذي اعتمدت عليه هذه الدلالات هو (الأنطولوجيا العُليا المدمجة المقترحة (**SUMO**)) وما يتّصل بها من أنطولوجيات مُختصّة بهذا المجال. وقد توسعت الدراسة بشكل كبير في الأنطولوجيا ومجموعة خرائطها لتوفير شروط اصطلاحية وتعريفات لكل مجموعة مرادفات. ومن الأدوات التي طوُّرت جزءاً من هذه الدراسة هي إنشاء منصّة معجميّة المعادلات على غرار تلك المستخدمة في إنشاء شبكة الكلمات الأوروبية (**Euro WordNet**)، مع تسهيلات إضافيّة للكتابة العربية. وفيما يأتي المباحث الرئيسية التي ناقشتها هذه الدراسة.

التّحديات

تختلف اللُغة العربية عن اللُغات الهندو- أوروبية من النّاحية التّركيبية، والصّرفية، والدلالية. ويدل مصطلح «العربية التقليدية أو الكلاسيكية» على الشّكل القياسي للُغة المستخدمة في جميع أعمال الكتابة واللُغة الكلامية المسموعة والمرئية سواءً على التّلفاز والمذياع أو في الخطابات العامّة والخُطب الدّينية. يحتوي نظام الكتابة باللغة العربية على خمسة وعشرين حرفا ساكنا وثلاثة أحرف (علَّة) متحركة طويلة ، وتتّخذ أشكالاً مختلفة وفقاً لموضعها في الكلمة. بالإضافة إلى حروف العلّة الطويلة هذه، تحتوي اللغة العربيّة على حروف العلة القصيرة، وحروف العلة القصيرة ليست جُزءاً من الحروف الهجائية بل تكون مكتوبة كحرف تشكيل فوق أو تحت حرف ساكن لإعطائه الصّوت المطلوب، وبالتّالي إعطاء الكلمة المعنى المطلوب. وتعد النصوص التي لا تحتوي على حروف علة أكثر ملاءمة من المجتمع الناطق بالعربية، فهذا هو الشكل المعتاد للمواد اليومية المكتوبة والمطبوعة (كالكتب والمجلات والصحف والرسائل، الخ). لكن عندما يتعلق الأمر بالقرآن الكريم ودواوين الشّعر المطبوعة والكُتب المدرسية وبعض القواميس العربية، تظهر حروف التشكيل بشكل جليّ. إلاّ أنه على الرغم من أن معظم الناطقين بالعربية يمكنهم قراءة النصوص مع أحرف العلَّة المشار إليها بوضوح، غير أنّ القليل منهم يمكنه كتابة النصوص باستخدام حروف التّشكيل الصّحيحة. وقد تكرست العديد من الجهود لمعالجة القواعد الصّرفية للغة العربية التي تظهر نتائجها في العديد من تقنيات التحليل والمولّدات الصرفية. وعلى الرغم من ذلك، من أجل إنتاج نظام قائم على أساس التحليل والتوليد الصرفي كفؤ لغويا وحسابياً، ينبغي أخذ العوامل التالية بعين الاعتبار:

عادة ما يرتبط نمط الكلمة مع عدد كبير من الجُذور.

يعتمد وجود شكل صرفي واحد على وجود أشكال أخرى مكونة من نفس الوحدة الصّرفية.

هناك بعض الحالات التي يكون فيها أكثر من وظيفة صرفية لشكل واحد من أشكال المفردة أو الكلمة.

يتم إنشاء الكلمة أو المفردة من خلال الجمع ما بين جذر مُشفّر يدوياً ونمط مُشكّل بحيث يلزم أن يكون كل منها مشفراً يدوياً (encoded manually) للإشارة إلى مجموعة فرعية من الأنماط (pattern diacritized) التي يمكن يرتبط بها الجذر.

يمكن استخراج الجذر عن طريق إزالة الزوائد لتحديد الشّكل الأساسي للكلمة المشكّلة، وتطبيقه على المقياس الصرفي أو النمط.

صُممت بعض التّقنيات بحيث لا تأخذ أي نص عربي كمُدخل مباشرة، ولكن بترجمة النظام العربي إلى (ASCII) ومن ثم تلقيمها للنظام. وبعد ذلك تُترجم النتائج للعربية مرة أخرى حتى تُفهم.

ويبدو أنه لا يوجد توافق -حتى الآن- حول أقرب طريقة للتحليل/ التوليد الصّرفي الكفُؤ، ولا توجد أية وسيلة مُناسبة لتوليد أو تحليل الجذور العربية بسبب تعقيد حروف العلة الضعيفة المهيمنة على عدد كبير من المفردات العربية.

الغموض المعجمي (lexical ambiguity)

قد يحمل العنصر المعجمي معنيين مختلفين وغير مترابطين (الجناس)، ويمكن تعريف الجناس بأنه كلمة لا توجد علاقة بين معانيها، ويُمثل الغُموض والبساطة في الصّيغ الاسمية إشكالية مُهمة تؤثر على ترتيب معنى الكلمة. ويختلف الغُموض ما بين لُغتين مختلفتَيْن عندما تقتبِس إحداهُما مُفردة من الأخرى، حيث يتّجِه تعدد المعاني أو الدلالات من لغة المصدر إلى لغة الاقتباس وليس العكس.

ولا أحد يجادل في مدى أهمية المعجم الدلالي في التعامل مع المعاني المختلفة ذات الصّلة بالكلمات والمفاهيم. لكن ينبغي أن يكون هناك اتفاق حول كيفية تمثيل البيانات المعجميّة بحيث يمكن معالجتها بسهولة من جهاز الحاسب الآلي من أجل ترميز أي علاقات دلاليّة بين المعاني وتنفيذ التطبيقات المعجمية المفاهيمية المختلفة، مثل التوضيح (إزالة الغموض)، والسّلاسل المعجمية وغيرها.

صناعة المُعجم

لاحقاً لصدور شبكة (Euro WordNet) ، طُوِّرت شبكة الحروف العربية (AWN) على مرحلتين: الأولى من خلال بناء «شبكة أساسية للكلمات» وتُعنى بالمفاهيم الأكثر أهميّة؛ وثانياً: توسيع نطاق الشبكة الأساسيّة نزولاً إلى مُستوى أدنى لتشمل مفاهيم أكثر تحديداً باستخدام معايير إضافية.

ومن هنا ينبغي أن تُصبح «الشبكة الأساسية» متوافقة إلى حد كبير مع شبكات الكلمات باللُغات الأخرى التي تُطَوَّر بنفس الأسلوب. بالنسبة لشبكة الكلمات الأساسية، شُفِّرت المفاهيم الأساسيّة المشتركة لـ ١٢ لغة كقوائم مرادفات في شبكة الكلمات العربية؛ كما أضيفت مفاهيم أخرى خاصة باللغة العربية وترجمتها يدوياً إلى أقرب معنى. وسيتم تنفيذ نفس الإجراء على جميع قوائم المرادفات الإنجليزية التي ترتبط حالياً بعلاقة تكافؤ في أنطولوجيا. (SUMO)

ويتخذ ترميز أو تشفير قوائم المرادفات اتجاهاً ثنائياً: فبالنّظر إلى قوائم المرادفات الإنجليزية، تُختار جميع المتغيرات العربيّة المقابلة (إن وجدت)؛ أما بالنّسبة للكلمة العربية، فسيتمّ تحديد جميع معانيها وسيتمّ (لكُل واحدة منها) تشفير قائمة المرادفات الإنجليزية المقابلة. أما قوائم المرادفات العربية، فيُوسع نطاق مُفرداتها الأعمّ (الأشمل) لتشكيل تسلسل هرمي دلالي مُغلق، وسوف تُستخدم أنطولوجيا (SUMO) لتعظيم الاتساق الدلالي لروابط الجناس (الخاص). وهذا ما يمثل الشبكة الأساسية التي تعتبر الأساس الدلالي لمزيد من الامتداد (التوسّع)، وأغلب العمل عليها يتمّ يدوياً.

الأدوات

وتشمل الأدوات التي سيتم تطويرها من أجل الشّبكة العربية (AWN) واجهة مُستخدم معجمية على غرار واجهة الشبكة الإنجليزية (EWN) مع تسهيلات إضافية للنّص العربي. وستكون واجهة المستخدم مُتعددة اللُغات، ولا تتأثر باتجاه المواءمة بين البنى المفاهيمية للغتين. وبالإضافة إلى تسهيلات البحث والتصفح المتاحة للمستخدمين النهائيين لقاعدة البيانات

المكتملة، تتطلب صناعة المعاجم واجهة للتّحرير، حيث تتوفر مجموعة متنوعة من المكونات التّراثية، ولكلٍ منها مزاياه الخاصّة. وسوف تكون واجهة المحرر متصلة مع خادم قاعدة البيانات بواسطة بروتوكول (SOAP)، مما يسمح لعدة معجميّن (مؤلّفي معاجم) على مواقع مختلفة بالاحتفاظ بقاعدة بيانات مشتركة.

قاعدة البيانات

تتكون هيكليّة قاعدة البيانات من أربعة أنواع من الهيئات (الكيانات) وهي: العنصر والكلمة والشّكل (الصّيغة) والرابطة (العلاقة) (link) ، حيث تمثّل العناصر الهيئات أو الكيانات المفاهيميّة (conceptual entities) بما فيها قوائم المرادفات، وفئات الأنطولوجيا (ontology classes) والأمثلة. أما كيان أو هيئة الكلمة فهو معنى الكلمة. في حين يدلّ الشكل (أو صيغة الكلمة) على الكيان أو الهيئة التي تتضمن معلومات معجميّة. وأخيراً، تُمثّل الرابطة (أو العلاقة) حلقة الوصل بين عنصرين، وتأتي على عدة أنواع مثل «التكافؤ» (equivalence) أو «التضمين»(subsuming) أو غيرها من الأنواع. وقد حُدد نموذج البيانات هذا على شكل تنسيق تبادلي (interchange format) بصيغة (XML)، لكنه يُنفذ أيضاً على قاعدة بيانات. MySQL

الأنطولوجيا

سيتم كذلك بناء أنطولوجيا ضخمة لتوفير الأساس الدلالي للمفاهيم التي تتضمنها الشبكة العربية (AWN) على أنطولوجيا (SUMO)، وهي عبارة عن أنطولوجيا اصطلاحية مكونة من حوالي ١٠٠٠ مصطلح و ٤٠٠٠ تعريف مُتاحة حاليا من خلال أنطولوجيا من الدرجة الأولى تُسمّى(الأنطولوجيا العُلوية القياسية لتبادل المعرفة (SUO-KIF)) وقد تُرجِمت أيضاً إلى لغة دلالية على الشّابكة (OWL). وتمتلك أنطولوجيا (SUMO) قوالب لتوليد اللغة الطّبيعية ومعجم متعدد اللغات يسمح بالتّعبير عن العبارات الموجودة في أنطولوجيا(SUO-KIF) و أنطولوجيا (SUMO) بلُغاتٍ متعدّدة.

الخاتمة

يؤدي بناء شبكة للنّص العربي إلى تحدّيات لا يمكن مواجهتها من خلال الشّبكات القائمة. وتشتمل هذه الشّبكة على النصّ (الخط) من جهة؛ والخصائص الصّرفية للغات السّامية التي تتمحور حول الجذور من جهة أخرى. وقد وُضعت الأسس اللاّزمة لمواجهة هذه التحديات. وتقترح الدراسة ابتكار شبكة لغوية ستكون ذات نتائج ملموسة من خلال إحلال أنطولوجيا (SUMO) محل شبكات الكلام الإنجليزية (مثل المؤشّر العابر للُغات ILI).

The Fourth Global Word Net Conference, Szeged, Hungary

Year of publication: 2008

Arabic WordNet: Current State and Future Extensions

Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M. Antonia Martí, William Black, Sabri Elkateb, James Kirk, Adam Pease, Piek Vossen, Christiane Fellbaum

شبكة الكلمة العربيّة "ووردنيت": الوضع الرّاهن والتوسعات المستقبليّة

مقدمة

نُقدم من خلال هذه الورقة استعراضاً للوضع الرّاهن لمشروع شبكة الكلمة العربية (Arabic Wordnet)، لا سيّما مُحتويات قاعدة بياناته، وصناعة المعجم، وواجهات المستخدم، ومتصفّح شبكة الكلمة العربيّة وعلاقته بالأنطولوجيا العليا المدمجة المقترحة (SUMO)، وراصد الكلمة العربية(Arabic word spotter) ، بالإضافة إلى التّقنيات شبه-التّلقائية لامتداد (توسع) (extension semiautomatic) لشبكة الكلمة العربية. وينصب التّركيز الرّئيسي لهذا العرض على الامتداد شبه-التّلقائي لشبكة الكلمة العربية باستخدام القواعد المعجميّة والصّرفيّة.

ما يزال مشروع شبكة الكلمة العربية قيد الإنشاء حالياً، وهو يسير وفق منهجيّة سبق أن طُوِّرت لمشروع مُماثل هو (EuroWordnet) . حيث اعتمدت المنهجية على مبدأ تعزيز التّوافق بين شبكات الكلمات والتركيز على التّرميز اليدوي لمجموعة من المفاهيم الأساسية، وأكثر المفاهيم أهمية وفقاً لمختلف الشّبكات. وكما هو الحال في شبكة (EuroWordnetهناك رسم دقيق للشّبكة العربية على (Princeton WordNet ٢,٠) ، وبالإضافة إلى هدفه المتمثّل في بناء شبكة الكلمة العربية، يهدف المشروع إلى توسيع المواصفات الاصطلاحية للمعاني والمرادفات باستخدام الأنطولوجيا العليا المدمجة المقترحة (SUMO) المستقلة لغوياً. وهذه المعلومات

مشتركة بين لغة ولغة، ويمكن أن تكون بمثابة أساس لتطوير أدوات حاسوبية تستند إلى الدلالة لتطبيقات المعالجة بين اللغات الطبيعية (crosslinguistic NLP applications[١]) لمختلف اللغات.

تضمّنت هذه الورقة مبحثين رئيسيّين: ناقش الأول الوضع الرّاهن لشبكة الكلمة العربية؛ وناقش المبحث الثاني التقنيات شبه التلقائية المختلفة لامتداد (توسع) شبكة الكلمة العربية.

المبحث الأول: الوضع الرّاهن لشبكة الكلمة العربية

أولاً: محتويات قاعدة البيانات

في الوقت الذي أُعِدَّ فيه هذا البحث، كانت الشّبكة تحتوي على ٩٢٢٨ قائمة مرادفات (منها ٦٢٥٢ اسمية، و٢٢٦٠ فعليّة، و٦٠٦ نعتية (صفات) وأخيراً ١٠٦ قائمة ظرفيّة). ويتضمّن هذا الرقم ١١٥٥ قائمة للمرادفات التي ترتبط بالكيانات (الهيئات) الاسمية، التي استُخرجت بشكل تلقائي (أوتوماتيكي) ودُقّقت بواسطة معجميّين مختصّين. ولأن هذه الأرقام تتغيّر باستمرار فقد خُصص رابط دائم التّحديث للمستخدمين.

ثانياً: واجِهات المستخدم

طُوِّرت واجهتان للمشروع تعملان على الشّابكة، الأولى واجهة المعجميّ المعتمدة على الشّابكة (Lexicographer's Web Interface)، التي صُممت لدعم مهمة إضافة أو تعديل أو نقل أو حذف المرادفات (synsets) على شبكة الكلمة. وتشمل الوظائف التالية:

إدراج المرادفات المُخَصّصة لكل مؤلّف (مُعجَمِي).

إدراج المرادفات باللُغة الانجليزية.

إدراج المرادفات حسب مُكافئاتِها.

إدراج المرادفات حسب تاريخ إنشائها.

إدراج المرادفات بدون المفردات المعجمية.

أما الواجهة الثانية فهي واجهة المستخدم المعتمدة على الشّابكة (User's Web Interface)، ووظيفتها تمكين المستخدم من استشارة شبكة الكلمة العربية والبحث عن الكلمات العربية، وجذور المفردات العربية ومرادفاتها والكلمات والمرادفات الإنجليزية الخاصة بشبكة (WordNet ٢,٠)، ويمكن تنقيح البحث من خلال اختيار الجزء المناسب من أقسام الكلام.

ثالثاً: علاقة شبكة الكلمة العربية بالأنطولوجيا العليا المدمجة المقترحة (SUMO)

تُشكل منهجيّة (SUMO) ونطاق عملها أكبر قاعدة اصطلاحية توصيفيه (أنطولوجية) متاحة للجمهور اليوم. وهي معترف بها رسمياً ولا تعتمد على تطبيق بعينه. وتحتوي المنهجيّة على ١٠٠٠ مصطلح و٤٠٠٠ مسلّمة (بديهيّة) و ٧٥٠ قاعدة، وهي المنهجيّة التّوصيفيّة الاصطلاحية الوحيدة التي رُسمت يدوياً لجميع قوائم المرادفات المستخدمة على الذخيرة (Princeton WordNet) وكذلك على الذخيرة (EuroWordNet)، وهذه إحدى الطرق التي يوفر وجود الأنطولوجيا الاصطلاحية من خلالها تبادل بين اللغات لا يقتصر على معجميّة أي لغة بشريّة بعينها.

رابعاً: متصفّح شبكة الكلمة العربية

إنّ متصفح شبكة الكلمة العربية هو تطبيق مستقل يمكن تشغيله على أي جهاز حاسب يحتوي على آلية جافا افتراضيّة. وفي الوضع الرّاهن ، تشمل خدماته الرئيسيّة: تصفح شبكة الكلمة العربية، والبحث عن المفاهيم في الشبكة، وتحديث الشبكة وفقاً لأحدث البيانات المستقاة من المعجميّين. ويمكن إجراء عمليات البحث باللغتين العربية أو الإنجليزية. وفي حين أنّه يصعب على المستخدمين غير المتمرسين على التعامل مع اللغة العربية معرفة كيفية تحويل الكلمة العربية التي نُسخت من صفحة الشبكة إلى نموذج الاقتباس المناسب، قمنا بدمج التّحليل الصّرفي العربي ضمن وظيفة البحث باستخدام نسخة من تطبيق (AraMorph).

خامساً: راصد شبكة الكلمة العربية

لقد طُوِّر راصد الكلمة العربية لغرض تزويد المستخدم بأداة لاختبار تغطية شبكة الكلمة العربية عن طريق تحديد تلك الكلمات الموجودة على صفحة أي شبكة عربية يمكن العثور عليها في شبكة الكلمة العربية. ويجري البحث عن الكلمات العربية أولاً في شبكة الكلمة العربية، وفي حال الإخفاق في العثور عليها، ينتقل البحث إلى بعض القواميس ثنائية اللغة. وتعتمد هذه الآلية على المحلّل اللغوي(AraMorph)، وبمجرد العثور على المكافئ اللغوي، يجري توفير الترجمة على مستوى الكلمة. كما يجري أيضاً توفير ترجمة الكلمات الشّائعة أو المستبعدة.

المبحث الثّاني: أساليب (التقنيات شبه-التلقائية) المختلفة لامتداد (توسع) شبكة الكلمة العربيّة

على الرغم من أن بناء شبكة الكلمة العربية كان يدوياً، فإنه قد بذلت بعض الجهود لأتمتة جزء من العملية باستخدام الموارد المعجمية ثنائية اللغة المتاحة. في حين أنّ استخدام الموارد المعجمية للبناء شبه التلقائي للشّبكات اللُغوية الأخرى غير الإنجليزية ليس بالأمر الجديد.

أما بالنسبة لشبكة الكلمة العربية، فقد قامت هذه الدراسة باستقصاء نهجين مختلفين محتملين: الأول إنتاج قوائم من الترجمات العربية المقترحة للكلمات المختلفة الواردة في قوائم المرادفات الإنجليزية وما يقابلها من مجموعة المفاهيم الأساسية. وفي هذه الحالة اعتُبرت مدخلات المهمّة المعجمية هي مجموعة المرادفات الإنجليزية وترجماتها العربية. من جهة أخرى، اشتُقت أشكال جديدة للكلمة العربية الموجودة فعلياً والتي بنيت يدوياً مع قوائم مرادفات الأفعال العربية باستخدام القواعد الإعرابية والاشتقاقية، ومن ثمّ إنتاج قائمة مقترحة من روابط المرادفات الإنجليزية لكل شكل. وفي هذه الحالة كانت المدخلات المستخدمة هي الفعل العربي، ومجموعة المشتقات الممكنة ومجموعة المرادفات الإنجليزية التي يمكن أن تكون مُرتبطة بالقوائم العربية المقابلة. وفي كلتا الحالتين، جرى التحقق من صحة قائمة المقترحات يدوياً من قبل المعجميّن.

الاستنتاجات والتّوصيات

استعرضت هذه الدراسة الوضع الحالي لمشروع شبكة الكلمة العربية، وناقشت التقنيات شبه التّلقائية لتوسيع نطاق التّغطية لهذه الشبكة. فمن ناحية، ناقشت الدراسة التّرجمات المقترحة المستندة إلى ثمانية من قواعد الاستدلال التي استُخدمت في شبكة (EuroWordNet)، ومن ناحية أخرى، وُصفت مجموعة من إجراءات التمديد شبه التلقائي لنطاق تغطية الشبكة باستخدام القواعد المعجميّة والصّرفيّة وقُدمت نتائج تقييمها الأولي. وتأمل هذه الدراسة أن يستمر العمل على توسيع نطاق قاعدة بيانات الشبكة سواءً بالوسائل اليدوية أو الوسائل التلقائية (الأوتوماتيكية) حتى بعد انتهاء المشروع الحالي. وأخيراً، تتطلع الدراسة إلى مجموعة واسعة من تطبيقات معالجة اللغات الطبيعية التي سوف تستفيد من استخدام هذا المورد القيّم.

1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), Sharjah, 2013, pp. 1-6.

Year of publication: 2013

Arabic WordNet semantic relations enrichment through morpho-lexical patterns

Mohamed Mahdi Boudabous, Nouha Chaâben, Fatiha Sadat

إثراء العلاقات الدلالية في "ووردنيت" العربية من خلال أنماط معجمية وصرفية

إن أنطولوجية (ontology) قاعدة بيانات المترادفات «ووردنيت» العربية هي أحد أهم الموارد المعجمية للغة العربية الحديثة، ولكنها تعاني من بعض الضعف بسبب غياب بعض الكلمات وبعض العلاقات الدلالية بين مجموعات المترادفات. يقترح هذا البحث طريقة لغوية تعتمد على أنماط معجمية وصرفية لإضافة علاقات دلالية لتحسين أداء «ووردنيت» العربية، التي تُستخدم في العديد من تطبيقات معالجة اللغة العربية.

«ووردنت» العربية هي النسخة العربية من «ووردنيت» الإنجليزية، وقد أغنيت أول نسخة منها يدويا لتحتوي في نسختها الحالية على ٢٣٤٩٦ كلمة عربية (أفعال وأسماء وصفات) وحوالي ١١٢٧٠ مجموعة مترادفات و١٨٥٢٢ علاقة.

تتألف قاعدة بيانات «ووردنيت» من أربعة أنواع هي مجموعات المترادفات، وكلمات (مصطلحات) وأنماط وروابط، وتكون مجموعات المترادفات عبارة عن مجموعة من الكلمات التي لها معنى مشترك، والمصطلح هو كلمة لها مغزى محدد، والنمط هو معلومات معجمية، أما الروابط فهي العلاقات بين مجموعات المترادفات.

لإثراء تغطية «ووردنيت» العربية للعلاقات الدلالية نقترح طريقة لغوية تعتمد على أنماط معجمية وصرفية، وتسمح هذه الطريقة بإيجاد علاقات دلالية جديدة بين مجموعات المترادفات باستخدام ويكيبيديا العربية، ويجري ذلك على مرحلتين، في الأولى تُعرّف أنماط معجمية وصرفية باستخدام قاعدة دراسة مستخلصة من ويكيبيديا العربية، وفي المرحلة الثانية نستخدم

هذه الأنماط لاستخلاص علاقات دلالية جديدة، ثم يجري التحقق من هذه العلاقات وإضافتها إلى «ووردنيت".

تجرى المرحلة الأولى على ثلاث خطوات، الأولى هي استخلاص أزواج مجموعات المترادفات، وتهدف هذه الخطوة إلى تعريف قاعدة بيانات من أسماء مرتبطة معا بعلاقات دلالية، ولذا نستخلص الأزواج المرتبطة بعلاقات دلالية ونصنفها بناء على نوع العلاقة، والخطوة الثانية هي بناء قاعدة الدراسة التي تعتمد على أزواج مجموعات المترادفات المستخلصة في الخطوة الأولى، والخطوة الثالثة هي تعريف الأنماط المعجمة الصرفية على يد خبراء لغويين.

أما المرحلة الثانية فتضم أربع خطوات، أولها هي إنشاء االذخيرة (Corpus) ويجري ذلك بالبحث عن الأسماء في «ووردنيت» العربية، ثم نجد التقاطع بينها وبين مقالات مقابلة لها في ويكيبيديا وتُنزّل، وفي الخطوة الثانية تجرى عملية ما قبل المعالجة للذخيرة، فتُقسّم المقالات إلى جمل وتُستخلص الجمل الدلالية وعلامات أقسام الكلام (part of speech tagging)، وفي الخطوة الثالثة نستخلص العلاقات الدلالية بعد إيجاد الأنماط المقابلة لكل جملة واعتمادا عليها، والخطوة الرابعة هي التحقق والإثراء، إذ تُحذف العلاقات المتكررة ويُتحقق من صحة العلاقات المتبقية على يد خبراء اللغة ومن ثم إضافتها إلى «ووردنيت".

بعد ذلك تُقيَّم هذه الطريقة بناء على نتائج خطوة التحقق التي يجريها الخبراء، ويشير التقييم إلى أن لبعض العلاقات نتائج جيدة ولأخرى نتائج غير مرضية، وهذا يعود لعدة أسباب هي اختلاف عدد الأنماط المحددة لكل علاقة، وأن حجم الذخيرة محدود، وإلى الأخطاء التي قد تحدث عند تقسيم المقالات إلى جمل، لأن ذلك يعتمد على علامات الترقيم، والسبب الأخير يعود إلى علامات أقسام الكلام التي قد تتحدد بشكل خاطئ أو تُعطي الكلمة أكثر من علامة.****

لبناء انطولوجيا معجمية للغة العربية ، يجب التأكد من النسق (pattern) بواسطة خبير مجال (domain expert) ثم تطبيق تلك الأنساق على بناء الأنطولوجيا. وقد اخترنا الويكبيديا العربية كذخيرة لغوية باتباع عدد من الخطوات لتحقيق هذا الهدف.

Int J Speech Technol (2016) 19:177–189

Year of publication: 2016

Towards an automatic extraction of synonyms for Quranic Arabic WordNet

Manal AlMaayah, Majdi Sawalha, Mohammad A. M. Abushariah

نحو استخلاص أوتوماتيكي للمرادفات "ووردنيت" القرآنية

نطور في هذا البحث نموذج استخلاص أوتوماتيكي يُستخدم لبناء «ووردنيت» عربية قرآنية تعتمد على القواميس العربية التقليدية. إن بناء مورد للكلمات القرآنية مثل «ووردنيت» العربية القرآنية يتطلب عدة موارد لغوية أخرى، فقد استخدمنا هنا ثلاثة قواميس عربية تقليدية، وتقوم منهجيتنا على ثلاث خطوات أساسية هي: ما قبل المعالجة للموارد ، ثم استخلاص مغازي الكلمات، ثم إنشاء المرادفات.

لاستخدام القواميس العربية التقليدية لا بد من معرفة طريقة ترتيب المفردات فيها ومعرفة الصرف واشتقاقات اللغة، وتحضير القاموس وهو أمر أساسي، فلا بد أن تكون كل كلمة مخزنة مع جميع معانيها المحتملة نصاً في سطر واحد، وأن يكون معها جذرها، وأن يُعطى كل فعل علامة القسم من الكلام. ثم أي كلمة يكون لها معاني عديدة تُخزن في «ووردنيت» العربية القرآنية. نستخدم لبناء نموذجنا عينة من الأصول النصية للقرآن (**Boundary-Annotated Quran (BAQ) Corpus**) التي تحتوي ٧٧٤٣٠ كلمة قرآنية، وهذه العينة هي جزء عم (الجزء الثلاثون) الذي نستخلص كلماته من الأصول النصية كاملة، ثم نحذف كلمات التوقف.

نهدف من بناء «ووردنيت» القرآنية إلى ربط كلمات القرآن بمرادفات لها في القرآن أيضاً، ولجمع عدد كبير من الكلمات والمفاهيم فإننا نجمع اشتقاقات جذور كلمات القرآن،كما أننا نحذف الإضافات المتصلة بالكلمات مثل حروف الجر والعطف والضمائر المتصلة لتسهيل العثور عليها في القاموس.

بعد انتهاء عملية ما قبل المعالجة، فإننا نخزن النتائج في قاعدة بيانات تتألف من ثلاثة جداول هي كلمات جزء عم وجذورها واشتقاقات الجذور، التي يُستخدم محتواها لاستخلاص معاني الكلمات التي ستكون مدخلات عملية استخلاص المرادفات وهذه الخطوة هي الطبقة الأولى من مشروعنا. نستخدم القواميس التي تمت معالجتها لربط اشتقاقات كلمات القرآن مع معانيها في عدة سياقات لفهم أفضل للكلمة، وتحضر معاني الكلمات لتكون مدخلات الطبقة الثانية التي خزنت أقسام الكلام والجذور والتعريفات لكل كلمة في جدول منفصل للأسماء والأفعال.

يتم جمع المترادفات في مجموعات مترادفات، وتحتوي كل مجموعة مترادفات على كلمة واحدة، أما عملية معالجة البيانات فتتمثل بعمليتي التقطيع (tokenization) وتكون بتقطيع النص إلى كلمات وعبارات، والاستخلاص (stemming) بهدف تقليل الاختلاف بين الأشكال الصرفية للكلمات ذات المعنى نفسه. في مرحلة المعالجة الحاسوبية تتم عملية فهرسة (indexing) لتنظيم موارد البيانات، فتُستخدم الفهرسة للحصول على الملفات (كل تعريف كلمة هو ملف) ذات الصلة، وتُستخدم الفهارس للحصول على المرادفات وأي معلومات متصلة بها.

وتم تقييم «ووردنيت» القرآنية من خلال ثلاثة مقاييس، هي الدقة، والاسترجاع، ومقياس ثالث يجمعهما معا، فتقاس الدقة باستخدام عدد الآيات المتصلة بالكلمة مع العدد الكلي من الآيات التي تم استرجاعها، أما الاسترجاع فيُقاس باستخدام عدد الآيات المتصلة بالكلمة التي تم استرجاعها مع عدد الآيات المتصلة بالكلمة، ونركز على المقياس الثاني لأنه يقيس أداء النموذج في أنظمة استعادة المعلومات، فنلاحظ أنه زاد استرجاع الملفات المتصلة من القرآن، وحصلنا على تحسن بنسبة ٢٧٪ في المقياس الثاني.

IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 2, January 2013

Year of publication: 2013

Using Arabic Wordnet for semantic indexation in information retrieval system

Mohammed Alaeddine Abderrahim, Mohammed El Amine Abderrahim, Mohammed Amine Chikh

استخدام ووردنيت العربية للفهرسة الدلالية في نظام استعادة المعلومات

في سياق أنظمة استعادة المعلومات العربية واسترشادا بالأنطولوجيا العربية ولتمكين هذه الأنظمة من الاستجابة بشكل أفضل لمتطلبات المستخدمين، فإن هذه الورقة تهدف لتقديم الملفات والاستعلامات بأفضل مفاهيم مستخلصة من ووردنيت العربية.

تُعرف الأنطولوجيات بأنها أدوات قادرة على معالجة المعرفة بما يتخطى المفاهيم، ويمكننا استخدامها في عدة مجالات مثل البحث عن المعلومات والترجمة التلقائية وغيرها، ويمكن استخدامها على عدة مستويات في أنظمة استعادة المعلومات، ونهدف هنا إلى معرفة تأثيرها في عملية فهرسة الملفات والاستعلامات فهرسة دلالية (semantic indexing). تهدف الفهرسة الدلالية إلى تصحيح أخطاء التطابق المعجمي (lexical matching) باستخدام الفهارس الدلالية بدلا من الكلمات المفتاحية البسيطة، وتهدف إلى الحصول على المعنى الصحيح للكلمة في النص من عدة معان محتملة للكلمة الموجودة في القواميس والأنطولوجيات.

نوضح هنا طريقة الفهرسة الدلالية بناء على ووردنيت العربية، فتبدأ هذه الطريقة بالحصول على مفاهيم ووردنييت من الملفات، فنصنع جدولا بالمصطلحات الموجودة في النص بعد استثناء كلمات التوقف، ثم نستخلص معاني هذه المفاهيم من مجموعات المترادفات (synsets) في ووردنيت العربية مع طريقة فك اللبس (disambiguation) التي تعتمد على احتساب المسافة بين معاني الكلمات، لتحديد معنى واحد صحيح (correct sense) لكل

مفهوم، أما للمصطلحات التي لا تنتمي لمفردات ووردنيت فإن النظام يستخلص جذرها أولا قبل المرور بالعملية السابقة.

استخدمنا في هذه التجربة ذخيرة لغوية (corpora) تتألف من ٢٢ ألف ملف عربي تغطي مجالات عدة. ووردنيت هي قاعدة بيانات معجمية تتبع مفهوم ووردنيت برنستون للغة الإنجليزية واللغات الأوروبية، وتتركب من مجموعات مترادفات ومؤشرات تصف علاقتها بغيرها من المجموعات، وقد تنتمي الكلمة فيها إلى مجموعات عدة وعدة فئات وتصنف هذه الفئات في أربع طبقات هي اسم وفعل وصفة اسمية وصفة فعلية.

لتقييم منهجية الفهرسة الدلالية فقد قسمنا التجربة إلى أربعة أنواع بحث:

بحث بسيط قبل الفهرسة الدلالية: استخدمنا ٧٠ استعلاما بسيطا كلمات مفتاحية مع فهرسة بسيطة للملفات.

بحث دلالي كلي: فهرسنا دلاليا قائمة من ٧٠ استعلاما ومجموعة الملفات المستخدمة للبحث.

تفكيك الاستعلام (Expansion of query): فهرسنا دلاليا قائمة من ٧٠ استعلاما فقط، واستخدمنا كلمة واحدة لفهرسة الملفات.

تمثيل دلالي للملفات: فهرسنا دلاليا قاعدة بيانات الملفات واستخدمنا ٧٠ استعلاما بسيطا كلمات مفتاحية.

إن مقارنة بسيطة للنتائج التي حصلنا عليها قبل استخدام منهجية الفهرسة الدلالية لتمثيل الملفات والاستعلامات وبعدها تمكننا من استنتاج أن هذه الطريقة تحسّن في أغلب الحالات عدد الملفات والملفات المتصلة المسترجعَة، أي أن الفهرسة الدلالية تحسّن الاسترجاع.

أما عند المقارنة بين طرق البحث الثلاث الأخيرة لمعرفة أفضل طريقة للفهرسة الدلالية من حيث الملفات التي نجدها والملفات المتصلة بها، فإننا نجد أن الطريقة الأولى هي الفضلى، أي أن الفهرسة الدلالية للملفات والاستعلامات معا تمثل أفضل نظام للبحث.

٣-٤ أبحاث التطبيقات

توزعت التطبيقات على موضوعات فرعية عددها ١١ موضوعاً، هي: التعرف الصوتي وتوليد الكلام، والقارئ الآلي، والترجمة الآلية، والبحث في النصوص، والتحليل الدلالي، والسؤال والجواب، وتلخيص النصوص، وتحليل الرأي، والتعرف على أسماء الأشياء، والتعليم والتعلم الآلي.

٣-٤-١ أبحاث التعرف الصوتي وتوليد الكلام

تضمنت أبحاث التعرف الصوتي وتوليد الكلام على ١٥ بحثا بينها ٤ أبحاث من نوع (أ)، هي: تمييز الكلام العربي غير المعتمد على المتحدث باستخدام آلة دعم المتجهات و التعرف الآلي على الكلام العربي المستمر غير المعتمد على المتكلم والمعتمد على ذخيرة لغوية صوتية متوازنة وثرية ، ونظام تركيب الكلام العربي باستخدام نظام ماركوف المخفي المسمى "HTS_ARAB_TALK"، ونظام التعرف على الكلام العربي في الزمن الحقيقي.

وتضمن – أيضاً- أحد عشر بحثاً من نوع (ب)، هي: دراسة مقارنة بين أداء الخلايا العصبية ونموذج ماركوف الخفي في أداء التعرف الآلي على الأرقام العربية المنطوقة، و تحويل الكلام إلى نص مكتوب باللغة العربية ونظام لتحويل النص العربي إلى كلام بناءً على الشبكات العصبية الاصطناعية ، ومقارنة تحليلية للخوارزميات (MFCC, DTW , ANN) للتعرف على الكلام العربي ، ومقدمة إلى نظام تمييز الكلام العربي باستخدام نظام CMUSphinx ، و فحص برامج التعرف على الأصوات العربية باستخدام نظام«CMU Sphinx والتحليل الإحصائي للأصوات العربية للتعرف على الكلام العربي المستمر، والتعرف الصوتي على الحروف الأبجدية العربية باستخدام الشبكات العصبية، ونحو قارئ قرآني متحكم به بالكلام، والتعرف على الكلام العربي متعدد الوسائط للعلاقة بين الإنسان والإنسان الآلي (الروبوت)، وتطبيقات تفاعلية، والتعامل مع الحروف المطبقة لتحسين أداء التعرف على الكلام العربي.

Agria Media 2011 and ICI-11 Conference and Exhibition on Information technology and Instruction technology, Eger-Hungary, pp 401-416, October 2011.

Year of publication: 2011

Speaker Independent Arabic Speech Recognition Using Support Vector Machine

Shady Y. EL-Mashed, Mohammed I. Sharway, Hala H. Zayed

تمييز الكلام العربي غير المعتمد على المتحدث باستخدام آلة دعم المتجهات

إن البحوث التي أجريت في حقل تمييز الكلام العربي قليلة مقارنة بغيرها من اللغات، وقد بلغت الدقة في أنظمة تمييز الكلام المعتمدة على المتحدث ١٠٠٪ ، إلا أن الدقة في الأنظمة غير المعتمدة على المتحدث ضعيفة.

تُعنى هذه الورقة البحثية بتمييز الكلام العربي غير المعتمد على المتحدث باستخدام آلة دعم المتجهات (SUPPORT VECTOR MACHINE)، ويُطبق النموذج المقترح على الأرقام العربية المنطوقة بشكل متصل باستخدام الشبكات العصبية كمثال، كما يمكننا تطبيق النظام على أي مجال آخر.

إن عملية تمييز الأرقام المنطوقة ضرورية في كثير من التطبيقات التي تستخدم الأرقام كمدخلات لإرسال المعلومات واسترجاعها، وقد أجري ذلك أولاً ببناء وحدة أساسية تتألف من ١٠٠٠ رقم مؤلف كل منها من ١٠ أرقام أحادية الخانة (digit) ، مسجلة من ٢٠ متحدثاً يختلفون في الجنس والعمر والحالة الجسدية وغير ذلك في وسط ذي ضجيج. وثانيًا: كل تسجيل حُوّل إلى ١٠ أرقام أحادية منفصلة. وأخيرًا استخدمت هذه الأرقام الأحادية لاستخلاص خصائصها باستخدام تقنيات MFCC) Mel Frequency Cepstral Coefficients) التي كانت مدخلات للشبكات العصبية (Neural Networks) لمرحلة التمييز، وبلغ أداء النظام ٩٤٪ عند استخدام آلة دعم المتجهات (SVM).

التعرف التلقائي على الكلام (Automatic Speech Recognition ASR) هو عملية تحويل إشارات الكلام المنطوق الملتقطة إلى سلسة من الكلمات المقابلة لها في النص.

يستخدم هذا النظام في العديد من التطبيقات، ويتكون نظام التمييز بشكل أساس من أربع مراحل: إشارات الكلام (Speech Signal)، ثم مرحلة ما قبل المعالجة (Preprocessing) ، ثم استخلاص الخصائصن (Feature Extraction) ، ثم التصنيف (Classification).

تشير عملية «معالجة إشارات الكلام" (Speech signal processing) إلى العمليات التي تجري على إشارات الكلام، أما «استخلاص الخصائص» فهو مصطلح تمييز النمط الذي يشير بدوره إلى قياس الخصائص (characterizing measurements) التي تجري على النمط أو الإشارة، وهذه الخصائص من المدخلات إلى المصنّف (classifier)هي التي تميز النمط.

تكمن صعوبة تمييز الكلام المنطوق في أمور عدّة:

متغيرات من المتحدث: قد تنطق الكلمة بطريقة مختلفة من شخص إلى آخر.

متغيرات من البيئة: البيئات الصوتية التي تستخدم فيها أنظمة التمييز قد يقاطعها عدة أمور مثل الضوضاء أو الصدى وغيرها.

اتصال الكلمات في الكلام الطبيعي: يأتي الكلام متصلًا -أي بلا فاصل بين الكلمة والأخرى - فيصعب تمييز الكلمات عن بعضها.

إن التحديات التي تواجه أنظمة تمييز الكلام العربي تكمن في وجود لغة فصيحة ولهجات عامية تختلف من مكان إلى آخر، ولكنها أيضاً تشترك في العديد من الخصائص على المستويين الصوتي واللغوي.

بعض التحديات التي تواجه أنظمة تمييز الكلام العربي، هي:

معرفة الكلمات: معنى الكلمات ضروري لتمييز المقصود من الكلام.

متغيرات الأنماط بسبب اختلاف اللهجات: ما يؤدي إلى اختلاف نطق الكلمة الواحدة رغم أنها نفس الكلمة كتابة.

التمييز الصوتي لمقطع صوتي قد يعتمد بشكل كبير على السياق الذي يصدر فيه الصوت، وذلك يدعى عادة (Co articulation)، أي أن الخصائص الصوتية للمقطع الصوتي متأثرة بالمقاطع المجاورة وبمكان المقطع الصوتي في الكلمة ومكان الكلمة في الجملة.

التشكيل: غياب التشكيل في الكلام العربي يؤدي إلى اختلافات كبيرة في لفظه.

النظام المقترح

بدأ النظام بتسجيل الأرقام العربية من عشرين متطوعاً ١٠ إناث و١٠ ذكور، ثم تستخدم تقنيات التقطيع (segmentation)شبه الأتوماتيكية والأتوماتيكية لتقطيع هذه الأرقام إلى الخانات الأحادية(digits)، ثم تطبق تقنيات استخلاص الخصائص لاستخلاص خصائص الخانات الأحادية. وفي نظامنا استخدمنا لذلك (MFCC)، وأخيراً تستخدم الشبكات العصبية لعمليتي التدريب (training) والاختبار (testing) لتمييز الكلام، وهنا نستخدم آلة دعم المتجهات (SVM) للتمييز.

نظام التسجيل

يتكون نظام التسجيل المستخدم من تبويبين (tab):

تبويب المسؤول: (admin) وفيه إعدادات الصوت التي تُستعمل لتعديل المتغيرات المستخدمة في تحويل الأرقام إلى أرقام أحادية الخانة، وفيه إعدادات البيانات التي تحتوي أسماء المتطوعين للاختيار بينهم عند التسجيل، والمعلومات المكونة من ١٠٠ رقم كل منها مكون من ١٠ أرقام أحادية الخانة متتالية.

تبويب المتطوعين: يسمح باختيار اسم المتطوع من القائمة، وبدء تسجيل الرقم المعروض ثم إنهاء التسجيل والاستماع له للتحقق منه، ثم تخزينه بشكل أوتوماتيكي في المكان نفسه، ويعطى اسمٌ يتكون من اسم المتطوع متبوعاً بالرقم الذي سُجّل والبيانات الخاصة به.

بنينا قاعدة بيانات تحتوي موجات الكلام لعشرين متحدثاً (كما وصفناهم سابقًا)،

فيسجل كل منهم ٥٠ عددًا، كل عدد يحتوي ١٠ أرقام أحادية الخانة متتابعة؛ فيصبح لدينا ١٠٠٠ ملف، ثم نقسم كل ملف أوتوماتيكياً إلى عشرة أرقام أحادية الخانة منفصلة؛ فيتكون لدينا ١٠٠٠٠ ملف (رقم أحادي الخانة).

صُمّم النظام لتدريب واختبار محركات تمييز الكلام الأوتوماتيكية، ليُستخدم في أنظمة تمييز المتحدث، والجنس، واللهجة، واللغة.

قسمنا هذه الأجزاء في الهيكل الأساسي إلى مجموعات منفصلة؛ الأولى للتدريب، وهي ٧٥٪ من قاعدة البيانات وتحتوي ما يقارب ٧٥٠٠ رقم أحادي الخانة مسجلاً، والثاني للاختبار ويشكل ٢٥٪ من قاعدة البيانات ويحتوي ٢٥٠٠ رقم أحادي الخانة مسجلاً.

نظام التقطيع (segmentation)

يؤدي تقطيع الكلام دورًا مهماً في تمييزه؛ إذ يقلل من الحاجة لذاكرة كبيرة ويقلل من تعقيد الحوسبة (computation) في أنظمة تمييز الكلام المتصل الكبيرة (vocabulary continues speech systems).

طبقنا تحويل فورير السريع (Fast Fourier Transform) على ملف الموجة الذي يمثل الرقم المسجل باستعمال حجم نافذة ملائم، ثم استعملنا مرشح نطاق الذبذبات (band pass filter) (من ٣٠٠ هيرتز إلى ٣٤٠٠ هيرتز) على الملف الناتج للتخلص من الضجيج في الإشارة، ثم طبقنا تحويل فورير السريع العكسي (IFFT) على الملف الناتج للحصول على الموجة الأصلية بعد التعديل.

ثم تطبق عملية التقطيع بتقنيتين: شبه أتوماتيكية وأتوماتيكية، في الأولى نتبنى معاملات التقطيع التي تحتوي عدة عوامل هي:حجم النافذة، والسعة الدنيا والتردد الأدنى والتردد الدني والتردد الأعلى والصمت الأدنى وأقل كلاماً، وأقل عدد كلمات، يدويًا عن طريق التجربة والخطأ.

أما في الثانية فإن المعاملات تُضبط أوتوماتيكياً للحصول على أداء أفضل باستخدام

K-Mean clustering. ويقدر بعملية التجميعclustering ، وهي تجميع مجموعة من الأنماط في مجموعات منفصلة.

نظام استخراج الخصائص

بعد تقطيع الكلام لا بدّ من استخراج الخصائص من الإشارة، وذلك ضروري لمعالجة البيانات المدخلة إلى الخوارزمية في حال كانت كبيرة جداً، لتحويلها إلى مجموعة خصائص عرض مصغرة؛ إذ إنها بيانات كثيرة تحتوي القليل من المعلومات. وإذا ما اختيرت الخصائص المستخرجة بعناية فمن المتوقع أن مجموعة الخصائص ستستخرج المعلومات المتصلة من البيانات المدخلة لتطبيق العملية المرغوبة على عرض مصغر بدلاً من المدخلات كاملة.

وقد استخدمنا في هذا النظام MFCC) Mel Frequency Cepstral Coefficients)

مصنّف الشبكة العصبية

استخدم في هذا النظام تقنيات الشبكة العصبية (Neural Network Techniques) وهناك نماذج عصبية عدّة كل منها له إيجابياتها وسلبياتها معتمدة على التطبيق، وبناء على تطبيقنا اخترنا آلة دعم المتجهات، وهي مجموعة من أساليب التعليم المشرَف عليها التي تحلل البيانات، وتميز الأنماط، وتستخدم للتصنيف وللتحليل التراجعي (regression analysis). وتطبق باستخدام kernelAdatron algorithm، الذي يشكل مستوى مرتفعاً hyperplane أو مجموعة منها، في مساحة ذات أبعاد لا نهائية أو كبيرة تُستخدم للتصنيف أو التراجع أو غيرها، وبشكل بديهي سيكون هناك فصل جيد من المستوى المرتفعhyperplane الذي يملك أكبر مسافة عن أقرب نقطة بيانات تدريب من أي فئة، فبشكل عام كلما كبر الهامش قلت أخطاء التعميم للمصنِّف.

أبسط طريقة لعرض نتائج التصنيف هي مصفوفة اللبس (confusion matrix) ، فهي تعرّف بتسمية التصنيف المرغوب على الصفوف، والتصنيفات المتوقعة على الأعمدة. وبما أننا نريد أن يكون التصنيف المتوقع هو التصنيف المرغوب نفسه، لذلك فإن الوضع المثالي أن نجد جميع الأمثلة في خانات قطرية على المصفوفة.

وقد أمكن استخدام هذا النظام باستخدام الشبكات العصبية عند تطبيقه على نطق بلهجة مصرية عامية في بيئة ذات ضجيج، فكان أداء النظام يقارب ٩٤٪ عندما استخدمت آلة دعم المتجهات.

The International Arab Journal of Information Technology, Vol. 9, No. 1, January 2012

Year of publication: 2012

Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus

Mohammad Abushariah, Raja Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman Khalifa

التعرف الآلي على الكلام العربي المستمر غير المعتمد على المتكلم والمعتمد على ذخيرة لغوية صوتية متوازنة وثرية

المقدمة

اللغة العربية من اللغات السامية ولها أهميّة كبيرة في العالم. وعلى الرغم من أهميتها، إلا أن الجهود البحثية حول التعرف على الكلام التلقائي باللغة العربية لا تزال للأسف غير كافية. وبما أن اللغة العربية الحديثة هي التي تدرس على نطاق واسع في المدارس والجامعات وتستخدم في وسائط الاتصال، وجب التركيز عليها في البحث العلمي وليس الاعتماد على اللهجات العامية.

التحليل الإحصائي ووصف النص والكلام

من أجل التعرف على الكلام باللغة العربية، لا بدّ من توفير مجموعة من التسجيلات الخطية التي يجب أن تكون غنية ومتوازنة تتصف بالسمة الغنية؛ بمعنى أنه يجب أن تحتوي على جميع صوتيات اللغة العربية، ويجب أن تكون متوازنة في الحفاظ على توزيع الفونيمات (phonemes) للغة العربية أيضا. ومن الضروري أيضا أن تستند هذه المجموعة من تسجيلات الكلام إلى مجموعة مكتوبة بعدد مناسب من الجمل والعبارات التي تُنشأ من الخبراء. ولذلك، من المهم اختيار مجموعة مكتوبة عالية الجودة من الجمل والعبارات قبل تسجيلها.

نص غني ومتوازن من الناحية الصوتية

يتطلب إنشاء نص غني ومتوازن من الناحية الصوتية اختيار مجموعة من الكلمات الغنية من الناحية الصوتية، وتجمع هذه الكلمات معا لإنتاج الجمل والعبارات، ويتم التحقق من هذه الجمل والعبارات والتحقق من التوزيع الصوتي المتوازن لها. قد تُحذف بعض هذه الجمل والعبارات أو تُستبدل من آخرين من أجل تحقيق توزيع لفظي كاف.

في عام ١٩٩٧ أنشأت مدينة الملك عبدالعزيز للعلوم التقنية قاعدة بيانات لأصوات اللغة العربية. والغرض من هذا العمل هو الوصول إلى أقل عدد من الكلمات العربية الغنية من الناحية الصوتية. ونتيجة لذلك، وأنشئت قائمة من ٦٦٣ كلمة غنية من الناحية الصوتية تحتوي على جميع الصوتيات (الفونيمات) العربية، التي تخضع لجميع القواعد الصوتية العربية هذا العمل هو العمود الفقري لتكوين الجمل الفردية والعبارات، التي يمكن استخدامها لتطبيقات تحويل النص إلى كلام. هذا وقد استند إنشاء القائمة من ٦٦٣ من الكلمات الغنية صوتيا على أساس الخصائص والمبادئ التوجيهية التالية:

١- تغطية جميع الأصوات العربية التي يجب أن تكون متوازنة .

٢-تحتوي على جميع القواعد الصوتية (phonetic rules) العربية، وهذا يعني تغطية جميع الأصوات العربية.

٣- وجود أقل عدد ممكن من الكلمات، بحيث لا تحتوي القائمة على كلمة واحدة يتحقق هدفها من وجود كلمة أخرى في نفس القائمة.

٤- أن تكون من الكلمات المستخدمة والمتداولة قدر الإمكان.

وجود نص لذخيرة لغوية(corpus) ، غني ومتوازن من الناحية الصوتية (الذخيرة اللغوية)، شرط مهم لتطوير أيّ نظام للتعرف التلقائي على الكلام. احتوى الجزء الأساسي المطور على تسجيلات من ٤١٥ جملة عربية، ووضعت ٣٦٧ جملة غنية ومتوازنة من الناحية الصوتية مما سبق أن طُوّرت في مدينة الملك عبدالعزيز للعلوم والتقنية، وسُجلت واستخدمت

لتدريب النموذج الصوتي، أنشئت ٤٨ جملة إضافية تمثل الأمثال العربية من مختص باللغة العربية لاختبار النموذج الصوتي. وقد بلغ عدد الناطقين باللغة العربية الذين سُجلّت أصواتهم ٢٠شخصا من دول عربية مختلفة تمثل ثلاث مناطق رئيسية في العالم العربي هي: (بلاد الشام والخليج وأفريقيا). سُجل هذا الكلام في استوديو مقاوم للصوت باستخدام برنامج سوند فورج (sound forge ٠٨) واستغرق العمل ما يقارب ثلاثة أشهر لاستكماله، ابتداء من مارس ٢٠٠٩ حتى يونيو ٢٠٠٩.

وقد جرى إثراء هذا الكلام مع أصناف من الناطقين باللغة العربية، مع الأخذ بعين الاعتبار الخصائص التالية التي تمثل:

- فئات عمرية مختلفة.
- جنسيات مختلفة.
- مناطق عربية مختلفة.
- مهن مختلفة.
- مؤهلات أكاديمية مختلفة.
- إتقان اللغة العربية.

تحضير البيانات الكلامية وتحليلها

يغطي هذا القسم جميع خطوات التحضير والتجهيز المسبق التي وضعناها من أجل إنتاج بيانات خطابية جاهزة للاستخدام، والتي تستخدم لاحقا لتدريب النموذج الصوتي، وتشمل ما يأتي:

- تجزئة الكلام العربي ذاتيا (Automatic Arabic Speech Segmentation) خلال جلسات التسجيل، حيث طلب من المتكلمين أن ينطقوا الجمل الـ ٤١٥ بالتتابع بدءا من الجمل التدريبية تليها أحكام الاختبار.

- حُفظت تسجيلات كل مكبر صوت في ملف منفصل، وأحيانا تصل الملفات إلى أربعة اعتمادا على عدد الجلسات التي يقضيها المتحدث لإنهاء تسجيل الجمل البالغ عددها ٤١٥ جملة. وتستغرق هذه الطريقة وقتا طويلا، لذلك، كانت هناك حاجة لتجزئة هذه الملفات.

- أجري التصنيف يدويا (manual classification) ومن ثم التحقق (validation) من صحة البيانات الكلامية.

- وضع تقسيم الكلمة العربية التلقائي لنواتج كل الملفات الممكنة في دليل واحد. لذلك، وضع دليل للمستخدم يبين تفاصيل تلك الملفات مع دلائل الجملة المقابلة. ونتيجة لذلك، كان يُنظر في الكلام الصحيح فقط لمزيد من الخطوات قبل المعالجة.

- يوضع لكل متكلم مجلد يحتوي على ثلاثة مجلدات فرعية، وهي: "جمل التدريب" و "جمل الاختبار" ومجلد "أخرى". يحتوي المجلد الفرعي "جمل التدريب" على ٣٦٧ مجلدا فرعيا، يمثل جمل التدريب الـ ٣٦٧، بينما يحتوي المجلد الفرعي "جمل الاختبار" على ٤٨ مجلدا فرعيا يمثلون ٤٨ اختبارا. يحتوي المجلد الفرعي على عبارات من الكلام. يحتوي كل مجلد فرعي على اثنين من المجلدات الفرعية الأخرى وهي "الصحيح" و "الخاطئ". والمصنفات المبوّبة تحت المجلد الفرعي "الصحيح" هي تلك المستخدمة في خطوات المعالجة اللاحقة. أما مجلد "أخرى" فيحوي ألفاظ المتكلم التي لا علاقة لها بالموضوع، ولم تدخل لا في التدريب ولا في الاختبار.

تضمنت التجربة السابقة ٤,٠٧ ساعة من بيانات التدريب الصوتية، واستند النموذج الصوتي إلى ١٦من توزيع خليط غاوسي (Gaussian mixture distributions) وعلى ١٦ من توزيع الحالات. (state distributions) ولقد حصل النظام على دقة تعرُّف على الكلمات بنسبة ٢٣,٩١٪ بوجود التشكيل، ونسبة ٥٤,٩٢٪ دون وجود التشكيل. أما بيانات التدريب الصوتية هنا فقد كانت ٧ ساعات ونتج عنها نسبة تعرف عند وجود التشكيل ٤٣,٩٣٪، وبدون تشكيل ٥٧,٩٤٪. .

وبالنسبة للمتكلمين المختلفين عند التلفظ بجمل متشابهة، فقد حصل النظام على دقة بلغت ٩٢,٩٥٪ و٢٩,٩٦٪ ونسب خطأ في الكلمات بمقدار ٧٨,٥٪ و ٤٥,٥٪ لحالتي

وجود تشكيل وعدم وجوده على التعاقب. أما عند التلفظ بجمل مختلفة من متكلمين مختلفين كانت نسبة التعرف ٠٨ ,٨٩٪ و ٢٣ ,٩٠٪ ونسب خطأ في الكلمات ٥٩ ,١٥٪ و ٤٤ ,١٤٪ مع وجود تشكيل وعدم وجوده.

الخلاصة

تقدم هذه الورقة تقريرا عن العمل الذي أجري لتطوير نظام عربي عالي الأداء يستند إلى نظام الكلام الغني والمتوازن من الناحية الصوتية. ويشمل هذا العمل إنشاء مجموعة من ملفات الكلام الغني والمتوازن من الناحية الصوتية مع التضمين الكامل لعلامات التشكيل (diacritical marks)، وبناء القاموس الصوتي العربي، ونموذج اللغة الإحصائية العربية. وتظهر نتائج هذا النظام العربي، الذي لا يعتمد على المتكلم، أنه يضاهي العديد من النتائج السابقة المعروفة في اللغة العربية أو قد يتفوق عليها.

3rd International Conference on Automation, Control, Engineering and Computer Science (ACECS'16), Proceedings of Engineering & Technology (PET), pp. 141-147

Year of publication: 2016

Arabic Speech Synthesis System using HMM: HTS_ARAB_TALK

Krichi Mohamed Khalil, Cherif Adnan

نظام تركيب الكلام العربي باستخدام نظام ماركوف المخفي المسمى: *HTS_ARAB_TALK*

تصف هذه الورقة البحثية نظام تركيب الكلام العربي باستخدام برمجيةHMM، ونظام التركيب الذي طورناه وسميناه (HTS_ARAB_TALK) وهو يستخدم الصوتيات العربية كوحدة تركيب لبرمجية HMM حيث استخدمت قاعدة بيانات عربية تدعى PADAS، والهدف الرئيسي لهذا النظام هو الحفاظ على تماسك النص الموحد الذي يترجم بالتركيز على برمجية HMM ، وفي تجاربنا عُرضت الخصائص الطيفية باستخدام معاملات ميل سيبسترام (Mel coefficients cepstrum)، أما للتحليل الموجي فقد استخدِم مرشح مستحث بالنبضات أو بالضوضاء .(noise or pulse excited corresponding MLSA). وبالإضافة إلى هذا البناء الأساسي وُظف نظام تركيب عالي الجودة(STRAIGHT)، وقد عمل النموذج الذي طُوِّر على تحسين تركيب النطق (speech synthesis) وعلى جودة بسيطة وواضحة في بيئة اللغة العربية.

يهدف نظام النطق هذا إلى تزويد المستخدمين بمخرجات منطوقة، يمكن استخدامها في عدة تطبيقات. ويمكن تقسيم طرائق تركيب النطق إلى أربع فئات: التركيب التلفظي (Articulatory synthesis)، وتركيب الصوت التشكيلي(formant synthesis) ، والتركيب التسلسلي(concatenative synthesis) ، والتركيب الإحصائي (Statistical Synthesizers.). وتتكون هذه الطرائق بشكل أساسي من جزئين: إجراءات اختيار وحدات

التركيب الأساسية وجزء التركيب وتدريبها، حيث تُستخدم المعلومات الصوتية لتوليد إشارة الكلام. إن إحدى الطرائق الواعدة هي استخدام نموذج الصوت المعتمد على السياق الممثل بنموذج ماركوف المخفي (hidden Markov model (HMM)). وباستخدام هذا النظام التجريبي نقارن بين طريقتي التركيب/ التحليل للنطق والتمثيل اللفظي:

تمثيل بسيط عن طريق معاملات ميل سيبسترام باستعمال برمجية SPTK toolkit

تمثيل أكثر تعقيدًا بطريقة التحليل/ التركيب عالية الدقة STRAIGHT

لقد استعمل عدد من أنظمة التركيب على اليابانية والإنجليزية وعدة لغات أخرى، أما نظام HTS_ARAB_TALK فهو أحدها وقد صمم خصيصًا للغة العربية.

نظام تركيب النطق باستعمال HMM

نشرت في عام ٢٠٠٧ مقالة تصف خصائص HTS لتقدم مجموعة من الأدوات المجانية تشكل نظام تركيب لفظي يعتمد علىHMMs. ومنذ ذلك الحين سيطر التركيب اللفظي المعتمد على HMMs على الأبحاث العلمية، فلهذه الطريقة عدة إيجابيات، فهي تستعمل عوامل وسيطة (parametric) ويمكن أن تُوظف معاملات HMMs لتغيير خصائص الصوت المولَد، وإذا تمت هذه التغييرات بحكمة بدقة فسيكون من الممكن تركيب عدة أنماط وخصائص صوتية من قاعدة بيانات صوت طبيعي واحدة، فصنع النماذج الإحصائية هو أمر ذاتي (أوتوماتيكي)؛ ولذا يكون تغيير الأنماط أسهل، ويمكن إضافة مركبة الزمن الحقيقي (real time component) لأن HMMs ملائمة للتغييرات ذات النمط المتحرك (الديناميكي).

لا بد من معرفة هذه المصطلحات:

وحدة الصوت (Phoneme): هي أصغر عنصر صوتي لا يمكن تجزئته، وهي حالة صوتية تُحدد طبيعتها من خصائصها المميزة.

HMMs : نموذج ماركوف المخفي (Hidden Markov models) تستخدم بصيغة الجمع والمفرد.

HTK : مجموعة من الأدوات للتعامل مع **HMMs** .

التدريب (**Training**): يشير إلى جميع العمليات لتشكيل وبنائع لفظ معتمد على **HMMs**.

التركيب (**Synthesis**): يشير إلى الإشارات الصوتية المركبة التي تستخدم النماذج المعتمدة على **HMMs**.

خطوات التركيب

يتم التركيب بخطوتين تتمان عن طريق **SPTK** أو **STRAIGHT**

الاستخلاص

كخطوة أولى لا بد من وجود معلومات كافية، أي أن تتوفر أمثلة كافية من كل مكون للتركيب، وفي حالة **HTS** هذه ستكون لفظا مكونا من الصوتيات.

من قاعدة البيانات هذه نجد أن الخصائص المستخلصة تتكون من نوعين: خصائص الإثارة (**excitation**) (خصائص متعلقة بالتردد الأساس لإشارة اللفظ في لحظة معينة) وخصائص طيفية (**spectral**).

من الضروري أن تكون قاعدة البيانات مشروحة (**annotated**)، وهي عملية تبين لكل ملف في قاعدة البيانات بداية ونهاية كل صوت في الملف. وبغياب نظام تمييز الأداء لا بـد أن يتم ذلك بطريقة يدوية، وذلك يتطلب وقتا طويلا. ستضاف معلومات أساسية لهذه الملاحظات ومدتها في ملفات النسخ، وستحتوي الملصقات على معلومات حـول الإشارة الصوتية مثل (المدة، الإشارة السابقة واللاحقة والحالية) ومعلومات عروضية (عدد الكلمات في الجملة وعدد المقاطع في الكلمة وموقع المقطع).

التدريب

تستخدم هذه المعلومات مرحلة تدريب **HMMs**. فبعد هذه الخطوة سيكون لكل سياق

معتمد ولكل لفظ HMMs نموذج ومدة، ومن المفهوم أن أهم قاعدة بيانات إحصائية ستكون واقعية أكثر وسيكون كل صوت ممثلا عدة مرات.

التركيب

في هذه المرحلة سيكون من الضروري تحديد النظام الذي نأمل تركيبه: لا بد من تحليل النص واستخلاص معلومات السياق اللفظية والعروضية من الملف المنسوخ المقابل للنص المطلوب، وتدعى هذه الخطوة التوليد، فهي ترجمة النص المكتوب إلى نسخ صوتية.

حالما يصبح الملصق جاهزا ستسمح المعلومات الموجودة فيه لـ HMMs بتوليد معاملات (المعاملات المناظرة وخصائص التنبيه المذكورة أعلاه).

في هذه النقطة تتوفر جميع المعلومات المطلوبة للتركيب، وسيتطلب الأمر توليد شكل موجي (يمكن أن يُسمع) من هذه الخصائص، وهذا يحدث من نموذج ترشيح موردي يكوّن الصوت البشري، وفي هذا النموذج يجذب المورد مدخلات المرشح الذي ستكون مخرجاته الشكل الموجي المتوقع. ومن ثم:

سيكون المورد إما سلسلة نبضات ديراك (Dirac pulse train) التي تمثل اهتزازات الأحبال الصوتية) أو ضجيجا أبيض (White noise) ناتجاً عن عدم اهتزاز الأحبال الصوتية).

أما المرشح، فقد ولد المعاملات الطيفية، وهو محاكاة لشكل المجرى الصوتي في وقت التوليد.

التحقق الأولي من العملية المتبعة

أ) دراسة قواعد البيانات

الوضع المثالي أن تتوفر قاعدة بيانات بكل الصوتيات العربية، والجزء الصوتي من قاعدة البيانات يتكون من ملفات .wav

توليد ملفات النسخ (الملصقات).

تكمل هذه الملصقات قاعدة البيانات فهي تحتوي جميع المعلومات التي تتيح للآلة معرفة محتويات الملفات الصوتية. وتُرفق مع قاعدة البيانات ملفات نصية (ملف نصي لكل ملف صوتي)، وهي تقدم معلومات حول بداية الوحدات الصوتية (phonemes) المكونة للملفات الصوتية ونهايتها والملفات النصية مولدة من برمجية تدعى speech processing Praat.

ملفات السؤال

هي ملفات نصية تحدد الأسئلة الموجهة إلى عُقد (nodes) شجرة القرارات (decision trees) لعناقيد HMM clustering. علماً بأن الأسئلة التي تُطرح على الصوتيات تتصل مباشرة بالمعلومات الأساسية التي تقدمها الملصقات.

تطوير نظام

يتكون النظام من ثلاثة مكونات أساسية: تدريب HTS (HTS –training) ومحرك HTS (HTS-engine)) ولوحة مفاتيح عربية، نحضر في أول جزء قاعدة بيانات عربية عروضية، ونقوم ببناء اللفظ الإحصائي للمعاملات الوسيطة، وبعدها نرسل المعاملات إلى الجزء الثاني، والنص هو مدخل النظام.

أ) تقسيم النص

سيقسم محلل المقاطع اللفظية النص إلى وحدات من مقاطع لفظية حسب القواعد العربية. يعتمد البناء على المدخلات وعملية المعالجة وتخطيط المخرجات، وسيحول هذا النموذج المدخلات الرمزية إلى نص مقروء، وقد يكون النص المُدخل فقرة أو جملة أو كلمة، لذلك لا بد من تقسيمه بترتيب هرمي، المستوى الأعلى للفقرات، ثم الفقرات إلى جمل، ثم الجمل إلى كلمات ثم الكلمات إلى مقاطع لفظية ثم المقاطع اللفظية إلى وحدات صوتية.

ب) توليد الشكل الموجي

HTS-engine-API : خدمين بناء برمجيتهم بناء على محرك التركيب، وفي الحقيقة هناك جزء من **HTS-engine** أُدخل في عدة برمجيات مثل **Open MARY**، **Festival** ، **ATR XIMERA**.

بعد ذلك جرى اختبار للنظام باستخدام صوت أنثوي وذكري، وفي هذا الاختبار استخدم **vSPTK** أو **STRAIGHT**مائتي جملة وكانت كمية بيانات التدريب ٣٩٨ جملة. واستخدم أيضا مورد نطق طبيعي فقام ١٥ شخصا بإجراء الاختبار ثم جرى تقييم الجودة وكانت النتيجة أن استخدام **STRAIGHT** كان أكثر فاعلية بالمقارنة مع **SPTK** .

International Journal of Computer Applications (0975 – 8887), Volume 81 – No.4, November 2013

Year of publication: 2013

Real-Time Arabic Speech Recognition

Zaid Y. Mohammed, AbdulSattar M. Khidhir

نظام التعرف على الكلام العربي في الزمن الحقيقي

نبذة مختصرة

يقصد بالزمن الحقيقي أن التعرف الصوتي يجري في وقت إنجاز الكلام (بالتزامن). لذلك يحتاج التعرف على الكلام إلى تنفيذ حسابات بالغة التعقيد في زمن قصير، وهذا يُعدُّ تحديا كبيرا لأنظمة الزمن الحقيقي. ومع ذلك فإن استخدام خوارزميات بسيطة وسريعة قد تفي بالغرض. إن الهدف الأساسي من هذا البحث هو تصميم نظام تعرّف على الكلام العربي باستخدام ماتلاب، و يتعرف هذا النظام على بعض الحروف بدقة مع الحفاظ على ميزة السهولة والسرعة، وهو يستخدم طريقة (Mel-frequency cepstral coefficients (MFCC)) كوسيلة لاستخلاص الميزات وطريقة المسافة الإقليدية لمقارنة الصوت المفحوص مع قاعدة المعلومات.

المقدمة

أصبحت عملية التعرف على الكلام أكثر شيوعا بسبب زيادة استخدام الأنظمة الرقمية المدمجة (digital embedded systems) مثل الحواسيب والهواتف والسيارات والألعاب وغيرها من الأجهزة. يجب على هذه الأنظمة أن تكون قادرة على فهم اللغة العربية، فالفكرة الأساسية هي تحويل هذه الخوارزميات إلى نص باستخدام الحاسوب في الزمن الحقيقي. هناك عدة خوارزميات لإنجاز هذه التحويلات التي تعتمد على كيفية معالجة الإشارات الصوتية من

ناحية كيفية استخلاص الميزات وكيفية التعرف على الكلام وتحديد هذه الميزات ومدى مناسبة سرعة هذه الخوارزميات لأنظمة الزمن الحقيقي.

إن طريقة **MFCC** هي خوارزمية جيدة جداً لتطبيق التعرف على الكلام المرتكز على (الإدراك السمعي للإنسان)، وهي مستخدمة في هذا البحث لاستخلاص الميزات (لكل حرف). وبسبب بساطة طريقة (مسافة إقليدس) فقد استخدمت لمطابقة الميزات وتحديدها. الهدف الرئيسي من هذا البحث هو تصميم نظام التعرف وتنفيذه على الكلام العربي غير المعتمد على المتكلم.

التعرف على الكلام – خوارزمية التعرف على الكلام

تلتقط لاقطة الصوت إشارات الكلام، وهذه الإشارات تُجمع كعينات وتُحول إلى شكل رقمي بواسطة المحول التناظري إلى رقمي (**A/D**) عند تردد١١٠٢٥هيرتز. وتتضمن الخطوات مرحلة التهيئة الأولية(**Pre-emphasis**)، والتأطير(**Framing**)، ووضع النوافذ **windowing**)) و(**MFCCs**)، ثم يجري التعرف بحساب أقل مسافة إقليدية بين الـ(**MFCCs** وقاعدة البيانات لتحديد الحرف الذي تم التلفظ به.

استخلاص الميزات

دقة التعرف على الكلام تعتمد بشكل كامل على الميزات التي استخلصت من إشارات الكلام الداخلة، ويحوي الاستخلاص ميزات أفضل مع أقل نسبة خطأ. وقد اختيرت خوارزمية الـ (**MFCCs**) لأنها أقل حساسية للاختلافات التي تعتمد على المتكلم والتي تظهر من إشارات الكلام، وهي مرتكزة على إدراك السمع عند الإنسان الذي هو خطي الفواصل (**linear spaced**) عند ترددات أقل من ١٠٠٠هيرتز ولوغاريتمي الفواصل عند ترددات أعلى من ١٠٠٠هيرتز.

الخطوة الأولى التهيئة الأولية:**Pre-emphasis** وهي تعديل الطيف **Spectrum**

normalization

عملية التهيئة هي تعديل إشارات طيف الكلام، لكي يقوم المرشِّح بتمرير الترددات العالية ذات الرتبة الأولى (**First order high pass filter**) لغرض تخفيض الطاقة العالية التي تحويها حزم الترددات المنخفضة.

الخطوة الثانية التأطير: (**Framing**)

من المفترض أن تكون إشارة الكلام إشارة ثابتة إذا لم تُقسّم إلى إطارات (**Frames**)، هذه الإطارات تحدد تعقيد النظام و كفاءته، ومن المفترض أن تعالج الإطارات ذات الحجم الصغير في فترة قصيرة وتنتج بيانات مسهبة (**redundancy data**)، بينما الإشارة المستقرة **signal stationary** قد تُستنزف في إطار بحجم كبير. حجم الإطار عادة يكون بين ١٠-٢٠**ms** مع وجود ٥٠٪ من التداخل (**overlapping**).

الخطوة الثالثة (**Hamming Windowing**):

عملية التأطير تنتج إطارات متقطعة، وتعمل كنافذة هامنج (**Hamming Windowing**) وتستخدم لتقليل هذه التقطعات بقدر الإمكان.

الخطوة الرابعة: تحويل فوريه السريع (**Fast Fourier Transform**)

يعطي نطاق التردد معلومات عن إشارة الكلام أكثر مما يعطيه نطاق الزمن من معلومات، وهكذا، فإن تحويل فوريه السريع يهدف إلى تحويل الإشارة من نطاق الزمن الى نطاق التردد. إن عملية الالتواء (**convolution**) في نطاق الزمن بين الحبال الصوتية ورنين المسالك الصوتية يمكن أن يحول إلى عملية ضرب في نطاق التردد لكي تفصل بـ(**Cepstral analysis**)، هذا الفصل ينتج تعرفا على الكلام غير معتمد على المتكلم.

(**MFCCs**)

لمحاكاة إدراك الإنسان يتم الانعطاف (**warping**) من التردد بوحدة الهرتز إلى مقياس ميل(**mel**) وفق معادلة خاصة.

يمكن أن ينجز الانعطاف باستخدام صفوف (banks) من المرشِّحات المثلثة الذي هو خطي السرعة تحت الـ ١٠٠٠هيرتز ولوغاريتمي السرعة فوق الـ ١٠٠٠ هيرتز، إذ تحتوي الترددات تحت الـ ١٠٠٠هيرتز على معلومات أكثر من غيرها من الترددات، لذلك تستخدم صفوف مرشحات مثلثة أكثر لالتقاط هذه المعلومات.

مطابقة الميزات

بعد معالجة الـ(MFCCs) تكون النتيجة متجهات ميزات ذات ٢٠ بعدا، وسوف تقارن متجهات الميزات هذه مع الوحدات المرجعية (قاعدة البيانات)، و تستخدم مسافة إقليدس لحساب المسافة بين متجه الميزات لحرف غير معروف اللفظ مع كل الحروف المخزنة في قاعدة البيانات.

بعد حساب كل المسافات الإقليدية بين متجهات الميزات للحروف غير المعروفة وتلك المخزنة في قاعدة البيانات التي تمثل جميع الحروف، فإن الحرف الذي يمتلك أقل مسافة اقليدية يُختار ليكون الحرف الملفوظ.

النتائج والمناقشة

نظام التعرف على الكلام الموجود في هذا البحث نُفذ وجرى العمل عليه في بيئة الماتلاب. وقد سُجلت إشارة الصوت من ٤ أشخاص: ٢ من الذكور و٢ من الإناث، هذه الإشارات تعبر خلال مكونات النظام وبعد عملية (المعالجة القبلية واستـخراج الميزات) يبدأ طور المقارنة، وتجرى هذه المقارنة بحساب وإيجادها أقل مسافة إقليدية للاستقصاء عن الحرف الملفوظ.

الاستنتاجات

في هذا البحث قمنا بتصميم نظام التعرف على الكلام العربي في بيئة الماتلاب التي تعالج مسبقاً (Pre-process) الإشارة مع التهيئة المسبقة (Pre-emphasis) والتأطير واتخاذ النوافذ

(windowing & Framing) واستخدمنا خوارزمية (MFCCs) لاستخلاص الخصائص من إشارات الكلام، واعتمدنا على مسافة إقليدس لمقارنة الصوت الاختباري وقاعدة البيانات، وقد كان معدل التعرف بين ٤ ,٨٨٪-٨٦ ,٩٠٪.

عملنا المستقبلي سوف يستخدم الـ(MFCCs) ومسافة اقليدس في تنفيذ نظام التعرف على الكلام المرتكز على (FPGAs).

JKAU: Eng. Sci., Vol. 19 No. 1, pp: 43-60 (2008 A.D. / 1429 A.H.)

Year of publication: 2008

Comparative Study of ANN and HMM to Arabic Digits Recognition Systems

Yousef Ajami Alotaibi

دراسة مقارنة بين أداء الخلايا العصبية ونموذج ماركوف الخفي في أداء التعرف الآلي على الأرقام العربية المنطوقة

أحد الاختلافات بين اللغة العربية وبعض اللغات مثل اللغة الإنجليزية هو طريقة لفظ الأرقام العشرة: من الصفر حتى رقم تسعة.

إن جميع الأرقام العربية متعددة المقطع – باستثناء لفظ الصفر الذي يُعدُّ من مقطع واحد، وتحتوي أغلبية أصوات نطق هذه الأرقام على أصوات فريدة؛ أي أصوات حلقية وأصوات مفخمة.

لقد أكملنا في هذه الورقة بحثنا بتصميم خوارزمية نموذج ماركوف الخفي الذي اختُبِر للتعرف التلقائي على الأرقام العربية. وكان النظام القديم يتخذ الكلمة كوحدة للتعرف والمعالجة، أما النظام الحالي فاعتمد الفونيم (أي الصوت) كوحدة بديلة. عُوملت الكلمة ونُطقت بمعزل عن الكلمات الأخرى.

تضمن البحث تطبيق النظامين بطريقتين؛ وفق العينات الصوتية للاختبار والتدريب. ففي الطريقة الأولى مُزج بين المتحدثين في مرحلتي التدرب والاختبار، أما الطريقة الثانية فكانت عينات المتكلمين في التدريب مختلفة عن تلك التي في الاختبار.

إن الهدف الأساسي من هذا البحث هو مقارنة أداء هذين النظامين (الخوارزميتين) وتحليلهما ومناقشتهما في عملية التعرف على الحروف والأصوات.

حقق نظام التعرف على الأرقام العربية باستعمال الشبكات العصبية الاصطناعية نسبة دقة

٥ ,٩٩٪ و ٩٤٪، بينما حققت خوارزمية ماركوف الخفي نسبة دقة ١ ,٩٨٪ و٨ ,٩٤٪

الإطار التجريبي:

إن قاعدة البيانات المستخدمة في تدريب واختبار الطريقتين هي نفسها، بالإضافة إلى ذلك، فإن مؤشرات القيم الشائعة المتعارف عليها في كلا النظامين قد ثبتت لتكون نفس القيم.

نظرة شاملة لنظام الشبكات العصبية الاصطناعية

تتكون منظومة الشبكات العصبية من عدة مراحل تبتدئ بنموذج معالجة الإشارات الرقمية الذي يختص بالحصول على النطق باستخدام الميكروفون، والترشيح وأخذ العينات. لقد استخدم مرشح إمرار نطاقي بترددات بين ١٠٠ هيرتز و ٤,٨ كيلو هيرتز لترشيح الإشارات المنطوقة (إشارات الكلام المنطوق) قبل معالجتها، وثُبِّت معدل العينة لـ ١٠ كيلو هيرتز وبوضوح ١٦ بت لكل الكلمات المنطوقة المسجلة. وقد استخدمت أيضا طريقة يدوية لإيجاد النهايات لفصل الكلام المنطوق عن الجانب الصامت للإشارة، ويكشف ذلك نقاط البداية والنهاية من الكلمة المنطوقة أيضا.

حُسِبت تقنيات الشفرة التنبؤية الخطية للأطر المتتالية لتكون ٦٤ نقطة (٤ ,٦ مللي ثانية). في كل حالة، استُخدِمت ٢٥٦ نقطة لنافذة هامينغ (ترميز خطي يستخدم لكشف وتصحيح أخطاء البيانات الرقمية وحفظها) لتحديد نقاط البيانات التي ستُحلل.

استُخدمت شبكة مستقبلات متعددة الطبقات (MLP) ذات تغذية استباقية متصلة بالكامل لتمييز الأرقام المنطوقة المجهولة. وتستخدم جميع الخلايا العصبية لشبكة اللوجستيات اللاخطية وخوارزمية التدريب على الانتشار العكسي .

تحتوي شبكة المستقبلات متعددة الطبقات على طبقتين مخفيتين، الطبقة المخفية الأولى ذات ٤٠ عقدة، والطبقة المخفية الثانية ذات ١٥ عقدة، وتتكون طبقة الخرج من ١٠ خلايا عصبية.

نظرة شاملة لنظام خوارزميات سلاسل ماركوف المخفية

طُورت آلة الاستقبال والإرسال الأتوماتيكية المستندة على نظام سلاسل ماركوف المخفية لتنفيذ أهداف هذا البحث، وقُسِّم هذا النظام إلى ثلاث وحدات.

طُبِّق النظام الثاني المقترح في هذا البحث باستخدام تقنية سلاسل ماركوف الخفية بمساعدة أدوات HTC. وصُممت آلة الاستقبال والإرسال الأتوماتيكية للكلام في البداية كمميز للكلمة على مستوى الصوت لثلاث حالات، المستمر، من اليسار إلى اليمين، مع عدم تخطي نماذج ماركوف المخفية.

أُنشئت قاعدة بيانات داخلية من العشرة الأرقام العربية، وطُلب من ١٧ من الذكور الناطقين باللغة العربية أن ينطق كل الأرقام عشر مرات.

ثم، تتكون قاعدة البيانات من ١٠ تكرارات من كل رقم ينتجها كل متكلم، بلغ مجموعها ١٧٠٠ رمز. وسُجلت جميع العينات الخاصة بمتكلم معين في جلسة واحدة.

خلال جلسة التسجيل، جرى تشغيل كل نطق مرة أخرى للتأكد من أن الرقم بأكمله ضُمِّن في الإشارة المسجلة، واستخدمت جميع الرموز المميزة البالغ عددها ١٧٠٠ رمزٍ لمراحل التدريب والاختبار اعتمادا على طريقة تشغيل النظام.

لدينا في هذا البحث شكلان من الأنماط، وهما وضع متعدد المتكلمين ووضع المتكلم المستقل. وقد استخدمت قاعدة البيانات هذه في النظامين بالطريقة نفسها في كلا الأسلوبين.

ACM Transactions on Asian Language Information Processing (TALIP) Volume 8 Issue 4, December 2009, Article No. 18, 18 pages.

Year of publication: *2009*

Automatic Speech-to-Text Transcription in Arabic

Lori Lamel, Abdelkhalek Messaoudi, Jean-Luc Gauvain Limsi-Cnrs

تحويل الكلام إلى نص مكتوب باللغة العربية

غالبا ما يكون من الضروري في اللغة العربية فهم معنى النص من أجل معرفة كيفية نطقه أو نطقه نطقاً صحيحاً. لمعالجة هذه المشكلة، يستخدم محلل باكوالتر العربي المورفولوجي (Buckwalter Arabic Morphological Analyzer) لاقتراح أشكال متعددة من الكلمات المفردة، وتُستخدام أداة التعرف على الكلام لتحديد الخيار الأنسب تلقائيا.

نظرة عامة على نظام التعرف

تشكل بيانات البث الإذاعي والتلفزيوني تحديا في نسخها إلى نص مكتوب لأنها غير متجانسة، وتحتوي على شرائح من مختلف الطبقات الصوتية واللغوية. وقد تكون الإشارة ذات جودة (استوديو) أو ربما تكون قد نُقلت عبر هاتف أو قناة صاخبة أخرى (أي تحوي ضوضاء إضافية وتشوهات غير خطية)، أو يمكن أن تحتوي على الكلام أو الموسيقى أو شرائح الموسيقى النقية. ويجري عادة إنتاج خطاب طائفة واسعة من خلال مكبرات الصوت مع أنماط مختلفة من الناطقين: مثل مقدمي أخبار يستضيفون برنامجا حواريا، والصحفيين في أماكن نائية، ومقابلات مع السياسيين وعامة الناس، ومتكلمين غير معروفين، ولهجات جديدة، وغير الناطقين باللغة العربية، وما إلى ذلك من أنماط الكلام. وربما توجد نفس العبارات المنطوقة من المتكلم نفسه بحيثيات ضوضاء مختلفة. وفي السنوات الأخيرة، انتقل تركيز البحوث من البيانات الإخبارية الإذاعية (التي أعدت في المقام الأول في ظروف الاستوديو) إلى نسخ ما يشار إليه بعبارة "الحوارات الإذاعية" (البرامج الحوارية والمناقشات والبرامج التفاعلية).

يتطلب هذا النوع من البيانات النمذجة الصريحة لآثار الكلام العفوية الأكثر شيوعا بكثير من الأخبار الإذاعية، وأيضا القدرة على التعامل مع الكلام من مجموعة متنوعة من اللهجات العربية. يجب إجراء حساب دقيق للنمذجة الصوتية لهذه البيانات المتنوعة.

الخلاصة

وصفت هذه الورقة التحسينات التدريجية على نظام النسخ الآلي للبيانات الإذاعية باللغة العربية، وتسليط الضوء على التقنيات المتقدمة للتعامل مع خصوصيات اللغة العربية. "وتتمثل إحدى التحديات في التدريب التعامل مع معلومات غير مكتملة، حيث إن معظم النصوص العربية مكتوبة دون علامات التشكيل، ومع ذلك فإن علامات التشكيل توفر معلومات مفيدة لنمذجة النطق ومعالجة المستوى الأعلى. وبعد الدراسات الأولية التي ركزت على بيانات الأخبار الإذاعية العربية الموحدة الحديثة باستخدام تمثيل صوتي ، استُكشفت طرق مختلفة لتقليل الاعتماد على البيانات الصوتية النقية والتعامل مع البيانات الأكثر تنوعا. يمكن اشتقاق العديد من أشكال الكلمات الصوتية باستخدام محلل باكوالتر المورفولوجي وتعديلاته. ومع ذلك، من الضروري أيضا أن تكون قادرة على توليد الكلمات المنطوقة للكلمات التي لا يستطيع محلل باكوالتر معالجتها. وقد اقترحت قواعد عامة لتوليد النطق مع حروف العلة، وقد استخدمت هذه الطريقة لتسهيل التدريب على البيانات غير الصوتية. وفيما يتعلق بنمذجة النطق، وضعت قواعد واضحة للتعامل مع المتغيرات ثنائية الجدلية المتكررة، فضلا عن الاختلافات المنهجية في اللغة. أدت النمذجة الصريحة ومحاولة إطالة فترة النطق وإدخال المتغيرات في النطق إلى تحسينات كبيرة في تحويل النطق إلى نص مكتوب.

Journal of Computer Science 5 (3): 207-213, 2009

Year of publication: 2009

An Arabic Text-To-Speech System Based on Artificial Neural Networks

Ghadeer Al-Said, Moussa Abdallah

نظام لتحويل النص العربي إلى كلام مستند إلى الشبكات العصبية الاصطناعية

نظام تحويل النص إلى كلام هو برنامج قائم على الحاسوب، يقوم فيه النظام بمعالجة النص ويقوم بقراءته بصوت عال. بالنسبة لمعظم التطبيقات، هناك طلب على هذه التكنولوجيا لتقديم نوعية جيدة ومقبولة من الكلام، وتُقيم جودة مركِّب الكلام من خلال تشابهه مع صوت الإنسان الطبيعي وسهولة فهمه ووضوحه. تركيب الكلام عالي الجودة يستخدم في مجموعة واسعة من التطبيقات في العديد من المجالات، مثل خدمات الاتصالات السلكية واللاسلكية، وتعليم اللغة، وتطبيقات الوسائط المتعددة، والمعونة لذوي الاحتياجات الخاصة.

يتكون نظام مركِّب الكلام من عنصرين رئيسيين، هما: وحدة معالجة النصوص، ووحدة معالجة الإشارات الرقمية. ولمكون معالجة النصوص مهمتان رئيسيتان: أولاً: يحول النص الخام الذي يحتوي على رموز مثل الأرقام والاختصارات إلى ما يعادل الكلمات المكتوبة، وغالباً ما تسمى هذه العملية تطبيع النص، ثم يحول النص إلى تمثيل آخر ويخرجه إلى المركب الذي يحول المعلومات الرمزية التي يتلقاها في الكلام.

التكنولوجيات الرئيسة لتوليد موجات الكلام الاصطناعية هي تركيب الشكل (**formant synthesis**) والتركيب المتسلسل (**concatenative synthesis**). كل تقنية لديها نقاط قوة ونقاط ضعف والاستخدامات المقصودة لنظام تركيبي عادة ما تحدد أي نهج يستخدم. ويعتمد مركب الكلام الذي بناه الباحثون في هذا العمل على نهج التركيب المتسلسل.

مع التقدم السريع في تكنولوجيا المعلومات والاتصالات، تمنح نظم الحاسوب

للمستخدمين على نحو متزايد الفرصة للتفاعل مع المعلومات من خلال الكلام، فالاهتمام بتركيب الكلام وبناء الأصوات آخذ في الازدياد. في جميع أنحاء العالم، طُورت مركبات الكلام للعديد من اللغات الشائعة مثل الإنجليزية والإسبانية والفرنسية، وأجري العديد من الأبحاث والتطورات على تلك اللغات، لكن لم تحظ اللغة العربية باهتمام كبير مقارنة باللغات الأخرى ذات الأهمية المماثلة، ولا يزال البحث باللغة العربية في مراحله الأولى. استناداً إلى هذه الأفكار، قدمت هذه الورقة نظاماً لتحويل النص العربي الذي استُخرج من محرك بحث إلى كلمات منطوقة. وقام الباحثون بتصميم نظام تحويل النص إلى كلام استخدموا فيه نهج تجميع الكلام المتسلسل لتركيب النص العربي. استند النظام إلى الشبكات العصبية الاصطناعية، وتحديداً نموذج التعلم غير الخاضع للرقابة(unsupervised learning). واستخدمت أحجام مختلفة من وحدات الكلام لإنتاج الكلام المنطوق، كما قام الباحثون ببناء قاموس من خمسمائة كلمة شائعة من اللغة العربية. اختيرت وحدات الكلام الصغيرة المستخدمة للتركيب لتحقيق مفردات غير محدودة من الكلام، في حين استخدمت وحدات الكلمة لتركيب مجموعة محدودة من الجمل. أظهر النظام دقة عالية في تركيب النص العربي وكان الكلام المنتج واضحاً للغاية. وخلصت الورقة إلى أن مركِّب النص إلى كلام قد بُني بقدرة على إنتاج عدد غير محدود من الكلمات بصوت عالي الجودة، ، ودقة عالية في تحويل النص المكتوب إلى كلام.

International Journal of Innovative Research in Advanced Engineering (IJIRAE), ISSN: 2349-2163 Volume 1 Issue 11 (November 2014)

Year of publication: 2014

Comparative Analysis of MFCC, DTW & ANN for Arabic Speech Recognition

Bidoor Noori Ishaq, Bharti W. Gawali

مقارنة تحليلية للخوارزميات (MFCC،DTW،ANN) للتعرف على الكلام العربي

لاقت استراتيجية التعرف على الكلام بشكل تلقائي عن طريق الكلمات المنطوقة، من شخص ما، اهتماما كبيرا من الباحثين. ويجري إنشاء كل كلمة منطوقة باستخدام تركيبة لفظية مكونة من مجموعة من حروف العلة والحروف الساكنة التي تكوّن إشارة الصوت. لقد قامت هذه الدراسة على البحث في هذا المجال بالاعتماد على بيانات عربية ونظام قائم بذاته مبني على خوارزمية معامل ميل للتردد (Mel-frequency Cepstral coefficient (MFCC))، وهي عبارة عن معاملات، تمثل الصوت استنادا إلى إدراك النظم السمعية البشرية، إضافة إلى خوارزمية انعطاف زمن المسافة (Distance Time Warping (DTW) وخوارزمية الشبكة العصبية(ANN) وقد جرى تدريب الشبكات العصبية على ثلاث عينات صوتية متنوعة سبق أن سُجلت في فترات مختلفة من الزمن، تكونت من عشرة متكلمين مختلفين يكررون نفس العبارة في أوقات مختلفة. أما بالنسبة للانعطاف الزمني في المسافة، فمحاذاة الوقت بطريقة مناسبة في الكلام هي واحدة من المشاكل الأساسية لقياس المسافة في عملية التعرف على الكلام. إن تحولا صغيرا في الزمن يؤدي غالبا إلى نتيجة غير صحيحة، فقد وجد أن هذه البرمجة تعتمد أساسا على البرمجة الديناميكية، وتستخدم هذه الخوارزمية لقياس التشابه بين سلسلتين زمنيتين قد تختلفان في الوقت أو السرعة.

صُممت قاعدة البيانات المستخدمة باستخدام مختبر الكلام المحوسب، و تتكون هذه القاعدة من حروف عربية، وأرقام، وكلمات وجمل. وقد بلغ حجم قاعدة البيانات هذه ١٥٩٠ عينة، مقسمة كالآتي: ٨٤٠ عينة من الأحرف، ٣٠٠ عينة من الأرقام، ٣٠٠ عينة من الكلمات، إضافة إلى ١٥٠ عينة من الجمل، وقد أمكن الحصول على هذه العينات بمختلف أشكالها عن طريق أخذ عينة مكونة من عشرة أشخاص من المتكلمين باللغة العربية كعينة للدراسة، ٣ من الإناث و ٧ من الذكور، وأمكن الحصول على مستوى دقة أفضل في التعرف بما يقارب ٩٠٪ مع نظام **MFCC**. وقد هدفت هذه الدراسة إلى مقارنة أداء هذا النظام مع غيره من الأنظمة التي استخدمت في هذا المجال. قورن نمط الاختبار مع النمط المرجعي للحصول على أفضل تطابق، وأعطت النتيجة التي توصل إليها البحث دوافع قوية لاستخدام هذا النظام للتعرف على الكلام في اللغة العربية، حيث أمكن الوصول إلى مستويات عالية من الدقة مع الحد الأدنى من الجهد والوقت. ومن الملاحظ أن هذه التقنية في تحسن مستمر في نظام التعرف على الكلام. وبناء على التجارب والاختبارات لهذا النظام فقد تبين أن التعرف على الأرقام يكون أسهل من التعرف على الكلمات.

Information and Communication Technologies International Symposium, ICTIS07 Fes

Year of publication: 2007

Introduction to Arabic Speech Recognition Using CMUSphinx System

H. Satori, M. Harti, N. Chenfour

مقدمة إلى نظام تمييز الكلام العربي باستخدام نظام CMUSphinx

نقدم طريقة لبناء نظام تمييز آلي للكلام العربي يعتمد على النظام مفتوح المصدر CMU Sphinx-٤, وهو نظام تمييز للكلام بمفردات كبيرة ومعتمد على المتحدث.

إن نظام تمييز الكلام هو تكنولوجيا تتيح للحاسوب تحديد الكلمات التي ينطقها الشخص، وقد طُورت العديد من الأنظمة منها نظام CMU Sphinx-٤الذي يعتمد على نماذج ماركوف الخفية. (Hidden Markov Models) وسنعمل على بناء نموذج يستخدم ميزات هذا النظام وسنعرض التكيف المحتمل لهذا النظام مع تمييز الكلام العربي.

صُمم Sphinx-٤ ليكون على درجة عالية من المرونة إذ يمكن استبدال الوحدات البرمجية (module) بوحدات أخرى، ووحدات التركيب الأساسية للنظام هي الواجهة الأمامية (frontend) وأداة فك التشفير(decoder) وقاعدة المعرفة(Linguist). تعمل الواجهة الأمامية على تحويل المدخلات (صوت مثلا) إلى سلسلة من ميزات المخرجات. أما قاعدة المعرفة فتقدم المعلومات التي تحتاجها أداة فك التشفير لأداء عملها، وهي مكونة من ثلاثة نماذج هي النموذج الصوتي الذي يتكون من تمثيل للأصوات يُنشأ بالتدريب باستخدام الكثير من البيانات الصوتية، والقاموس المسؤول عن تحديد كيفية نطق الكلمات، ونموذج اللغة الذي يحتوي تمثيلا لاحتمالية حدوث الكلمات.

أداة فك التشفير هي أهم مكون للنظام، فهي تقرأ الميزات من الواجهة الأمامية وتجمعها

مع بيانات من قاعدة المعرفة وتغذية راجعة من التطبيـق، ثم تنفذ بحثا لتحديد تسلسل الكلمات ذات الاحتمالية الأعلى، ويمكن تمثيل البحث بسلسلة من الخصائص.

إن نظام تمييز الكلام العربي الأوتوماتيكي يشبه نظام**Sphinx-٤**؛ إذ يستخدم ثلاثة أنواع من النماذج المعتمدة على اللغة، هي النموذج الصوتي الذي يمثل إحصائيا مدىً من التمثيلات الصوتية المحتملة للمقطع الصوتي، وقاموس اللفظ الذي يحدد كيفية نطق كل كلمة، ونموذج اللغة أو نموذج القواعد الذي يمثل أنماطا لاستخدامات الكلمة، وكل كلمة في هذا النموذج يجب أن تكون في قاموس اللفظ.

أُنشئ تطبيق يسمى (**Hello_Arabic_Digits**) وعُدلت العناصر الثلاثة لتلائمه.

أُنشئت ذخيرة لغوية (**corpus**) من الأرقام العربية العشرة، فقد اختير ستة أشخاص من الرجال وطُلب منهم أن ينطقوا الأرقام خمس مرات، وبالتالي تتكون الأصول النصية من ٣٠٠ مقطع صوتي.

لتقييم أداء هذا التطبيق، طبقنا بعض التجارب على عدة أفراد (ثلاثة رجال) طلب من كل واحدٍ منهم نطق عشرة أرقام عربية، وسُجل عدد الكلمات التي مُيزت بشكل صحيح ومن ثم احتسب متوسط نسبة التمييز لكل شخص.

كانت النتائج مُرضية للغاية مع الأخذ بعين الاعتبار الحجم الصغير جدا لأصول البيانات المستخدمة.

The International Arab Journal of Information Technology, Vol. 6, No. 2, April 2009

Year of publication: 2009

Investigation Arabic Speech Recognition Using CMU Sphinx System

Hassan Satori, Hussein Hiyassat, Mostafa Harti, Noureddine Chenfour

فحص برامج التعرف على الأصوات العربية باستخدام نظام"UMC "xnihpS

في هذا البحث نوقشت إمكانية التعرف على الكلام العربي(ASR)، وهي عبارة عن تقنية تسمح للحاسوب بتحديد الكلمات التي يتحدث بها شخص ما بالميكروفون أو الهاتف. ولهذه التقنية مجالات واسعة من التطبيقات، من أهمها: التعرف على الأوامر ، إملاء الكلمات ، الاستجابة الصوتية التفاعلية. ويمكن استخدامها لتعليم لغة أجنبية أيضا. كما يمكن أن تساعد المعاقين على التفاعل مع المجتمع، وهذا يجعل الحياة أسهل وأبسط.

اقترح الباحثون نهجا جديدا لبناء نظام التعرف الآلي على الكلام العربي باستخدام البيئة العربية، وقد دُرب هذا النظام على أساس الأداة مفتوحة المصدر المسماة بـ"CMU Sphinx-٤"، باستخدام الحروف العربية. وتتكون عملية التدريب من: تحويل البيانات الصوتية أي سيل الأصوات إلى سلسلة من حاملات المزايا (feature vectors)، وتحويل النص إلى سلسلة خطية من ثلاثية ماركوف (HMM) باستخدام قاموس اللفظ، والعثور على أفضل تسلسل لها. وقد ركز الباحثون اهتمـــامهم على (HMM) ما يُعدُّ نموذجا إحصائيا، حيث يفترض أن النظام الذي تجري نمذجته هو تطبيق لعملية ماركوف مع متغيرات غير معروفة، والتحدي هو تحديد المتغيرات المخفية، ومن ثم يمكن استخدامها لإجراء مزيد من التحليل.

من أجل الحصول على تقييم جيد لأداء التطبيق، أجرى الباحثون بعض التجارب على

مجموعة أفراد بلغ عددهم ٦٠ متحدثا من دولة المغرب (٣٥ من الذكور و ٢٥ من الإناث). وطُلب من كل واحد منهم أن ينطق ١٠ أرقام عربية ويكررها ٥ مرات وجرى تسجيلها، وبالتالي، تكونت المجموعة من ٣٠٠٠ رمز. ، أعيد تشغيل كل لفظ مرة أخرى خلال جلسة التسجيل للتأكد من أن الرقم بأكمله جرى تضمينه في الإشارة المسجلة، واستخدمت جميع التسجيلات الـ ٣٠٠٠ (١٠ أرقام، ٥ تكرارات، ٦٠ شخصا) لمراحل التدريب. ثم سُجل عدد الكلمات التي تعرف عليها النظام بشكل صحيح، ومن ثم حساب معدل التعرف لكل اختبار. وقد اعتمد على تضمين النظام **CMU Sphinx** في نمذجة العربية، وهو يتكون من النماذج الصوتية واللغة المتولدة والمدربة مع البيانات العربية. ولضبط متغيرات هذا النظام استخدم الباحثون برامج نصية جديدة في اللغة العربية، وفعلا عدلت وضبطت هذه المتغيرات بنجاح. بعد إجراء عدد من التجارب والاختبارات أمكن الحصول على نتائج مرضية جدا مع الأخذ بعين الاعتبار صغر حجم مجموعة التدريب التي استخدمت في العربية إذا ما قورنت مع المجموعات المستخدمة في اللغة الإنجليزية. وقد لوحظ أنه من أجل الوصول إلى أداء جيد في التعرف، من المستحسن تدريب النظام على مجموعات كبيرة (أكثر من ٥٠٠ صوت مختلف). لكن هنا لم تستخدم مجموعة بهذا الحجم، فالهدف الرئيسي في هذه الدراسة هو إظهار إمكانية تكيف النظام مع بيئة اللغة العربية.

International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 01– Issue 02, November 2012

Year of publication: 2012

Statistical Analysis of Arabic Phonemes for Continuous Arabic Speech Recognition

,Khalid M.O Nahar, Moustafa Elshafei, Wasfi G. Al-Khatib
Husni Al-Muhtaseb

التحليل الإحصائي للأصوات العربية للتعرف على الكلام العربي المستمر

المقدمة

التعرف التلقائي على الكلام يعرف بأنه عملية تحويل الموجات الصوتية (الإشارات الصوتية) إلى ما يقابلها من الكلمات أو الوحدات اللغوية الأخرى استنادا إلى خوارزمية محددة.

في التعرف على الكلام، من السهل أن ندرك الكلمات المعزولة، ولكن التحدي يكمن في التعرف على الكلام المستمر(**continuous speech**). إن إحدى الوحدات الأساسية في أي نظام تعرف على الكلام هي النماذج الصوتية (**acoustic models**). هذه النماذج الصوتية تحتاج إلى تمثيل مناسب من أجل التعرف عليها في وقت لاحق.

معظم أعمال البحث التي أجريت على التعرف على الكلام العربي المستمر تستخدم نماذج ماركوف المخفية لتحقيق معدلات مختلفة من دقة التعرف. تقاس دقة التعرف على الصوتيات بنسبة الصوتيات المتعرف عليها بشكل صحيح. وتتأثر الدقة بعدة عوامل كوجود ضوضاء؛ ومجموعة الفونيم المستخدمة؛ وعدد الحالات المخصصة لكل صوت، ومدة كل صوت.

معالجة البيانات واستخراج المعلومات (**Data Processing and Information Extraction**):

لغرض استخراج معلومات مفيدة من الملفات التي تُنشأ مسبقا، استخدم برنامج الماتلاب لحساب الترددات والطول والقيم المتوسطة والانحراف المعياري للطول وتوزيع الوسائل وتوزيع أطوال كل صوت، والوضع (القيمة الأكثر شيوعا في مجموعة من القيم) لطول كل صوت ومتوسط (القيمة المتوسطة في مجموعة من القيم) للأطوال. علاوة على ذلك، استخراج احتمال حدوث كل صوت.

في دراسات سابقة جرى تقديم تقنية لتحديد إشارة الكلام النقي في بيئة صاخبة وبعض أساليب التحليل الإحصائي للصوتيات، وأمكن الحصول على دقة عالية نسبيا من عزلة الصوتيات في بيئة صاخبة وتحسين النوعية. إن الدراسة الإحصائية وتحليل الصوتيات العربية ضرورية لتحسين أداء الأنظمة العربية الحالية، فمعرفة طول صوت معين يمكن استخدامها للحد من طول السلسلة التي تمثله، وهذا بدوره يزيد من سرعة التعرف ودقته. إن تجميع الصوتيات على أساس متوسط أطوال كل واحد منها يمكن أن يساعد على تضييق البحث عن الصوت المناسب خلال مرحلة التعرف، وهذا يزيد من السرعة والدقة. ويمكن أيضا استخدام التحليل الإحصائي لتطوير أساليب أخرى دقيقة لتجزئة الصوتيات.

يمكن استخدام معلومات نموذج اللغة، مثل بيغرام (Bigram)، لاستبدال صوتيات متعددة المحاذاة بالوضع الصحيح، وبالتالي تقليل معدل الخطأ في الكلمات. هناك حاجة إلى مزيد من التحقق في المستقبل للتأكد من تأثير التحليل الإحصائي السابق على معدل التعرف وتحديد أفضل تكوين لبناء النموذج الصوتي (acousticmodel). ومن الضروري أيضا إجراء هذا التحليل على ذخائر عربية أخرى ومقارنة نتائجها.

International Journal of Electric & Computer Sciences IJECS-IJENS Vol: 11 No: 01

Year of publication: 2011

Phonetic Recognition of Arabic Alphabet letters using Neural Networks

Moaz Abdulfattah Ahmad, Rasheed M. El awady

التعرف الصوتي على الحروف الأبجدية العربية باستخدام الشبكات العصبية

التركيز وجمع البيانات

خضعت عملية جمع التسجيلات الصوتية المستعملة في البحث لما يأتي:

كان استقبال الحروف الأبجدية العربية الصوتية (٢٨ حرفا) بواسطة ميكروفون حساس واسع النطاق، ثم حفظها كعينات من الإشارات الصوتية التي سيتم تحليلها (٢٠١٦ هو عدد ملفات الموجات الصوتية). وكانت التسجيلات لـ (٦) أشخاص مختلفين (٣-ذكور + ٣ إناث) مع ٣ محاولات لكل شخص وبأربع حالات تشكيل: الفتحة والضمة والكسرة والسكون. واستخدم مسجل الصوت وبرنامج **Wave_lab** لالتقاط إشارة الصوت، في حين استخدم ماتلاب لعملية التحليل والتعرف، ثم خزنت البيانات في شكل موجات وجرى تحليلها والتعرف عليها باستخدام برنامج ماتلاب.

الأسلوب المقترح

احتوت الخوارزمية المستعملة على أربع خطوات: (أ) إزالة الضوضاء وإزالة الصمت (ب) ميزة استخراج (ج) تحليل المكونات الرئيسية (د) الشبكة العصبية الاصطناعية.

الحد من الضوضاء وإزالة الصمت (**Noise Reduction and Silence Removal**)

إن الهدف من هذه الخطوة هو تحسين جودة الكلام باستخدام خوارزميات مختلفة، وذلك بتحسين الوضوح و / أو الجودة الإدراكية الشاملة لإشارة الكلام المتدهورة باستخدام تقنيات معالجة الإشارات السمعية. ويُعدّ تعزيز الكلام المتدهور بالضوضاء أو الحد من الضوضاء أهم مجال لتعزيز الكلام، ويستخدم في العديد من التطبيقات مثل الهواتف المحمولة، وأجهزة الصوت عبر بروتوكول الإنترنت، ونظم المؤتمرات عن بعد، والتعرف على الكلام، والوسائل السمعية .

تحليل المكونات الرئيسية (Principal Component Analysis)

هذه تقنية مشهورة لاستخراج الميزات ولتقليل الأبعاد، ويقوم التحليل فيها على افتراض أن معظم المعلومات عن الطبقات واردة في الاتجاهات التي على أساسها يكون الاختلاف أكبر ما يمكن.

كان استعمال شبكة عصبية اصطناعية (ANN) (Artificial Neural Network) يشمل الخطوات الأساسية التالية:

تصميم الشبكة العصبية مع عدد من المدخلات (متجهات تمثل كل نمط).

التحقق من مدى دقة الناتج الفعلي الناتج من إدخال محدد ومدى مطابقته لما هو مطلوب.

إنشاء الشبكة (Creation of the network)

استعملت بيانات التدريب التي كانت بشكل مصفوفة مدخلات تمثل جميع ملفات التدريب البالغة (٨٩٦) ملفا، وتمثل هذه الملفات محاولتين لأربعة أشخاص؛ وتمثل كل تجربة معاملات (principal component analysis (PCA)) من ١١٢ لفظ لـ ٢٨ حرفاً هي حروف الأبجدية العربية، وبالتالي فإن عدد الأعمدة يساوي ٨٩٦ وعدد الصفوف كان ٤٨ حيث يحمل ذلك حوالي ٩٣٪ من إجمالي الطاقة من ميزات كل حرف.

شبكة المحاكاة والاختبار (Network Simulation and testing)

وضعت شبكة محاكاة مدربة وجرى اختبارها من متحدثين اثنين، وصل أداء أخطاء شبكة المحاكاة إلى حوالي ٠,٠٤٥ بعد ٤٠ حقبة وحوالي ٠,٤٤٦ بعد ١٠٠ حقبة. وكُشف عن نتائج الإخراج من المصفوفة وجرى تسجيلها. وُجد أن عدد الأخطاء في الحركات كان الأكبر في الفتحة (٢٥) ثم السكون (١٧) ثم الكسرة (١٥) ثم الضمة (١٣). ولم يكن هناك أخطاء في حرفي الهاء والدال بينما كان أكبر عدد من الأخطاء في حرف الألف المهموزة (فوق وتحت) بنسبة ١٢,٥٪ يليها حرف التاء ٨,٩٢٪.

(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 11

Year of publication: 2011

Towards Quranic reader controlled by speech

Yacine Yekache, Yekhlef Mekelleche, Belkacem Kouninef

نـحو قـارئ قرآنـي متحـكم بـه بالكـلام

التعرف التلقائي على الكلام هو عبارة عن التكنولوجيا التي تسمح بالاتصال مع الآلات باستخدام الكلام. هناك العديد من التطبيقات التي تستخدم هذه التكنولوجيا مثل عمليات الأيدي الحرة (hands free operation) كالموجودة في السيارات أو المستخدمة من الأشخاص ذوي الإعاقة، وأنظمة المعلومات الحكومية، والرد التلقائي على الاستعلام (automatic query answering)، والاتصالات الهاتفية مع نظم المعلومات وغيرها.

المنهج الأكثر استخداماً في نظام التعرف التلقائي على الكلام هو المنهج الإحصائي لنموذج ماركوف المخفي (HMM) الذي يُدرب باستخدام منظومة تحتوي على موارد كلامية من عدد كبير من المتكلمين بهدف الحصول على أداء مقبول، ولسوء الحظ فإن هناك نقصا في هذه المنظومة للغة العربية. في هذه الورقة البحثية جمع الباحثون منظومة جديدة أسموها القيادة والسيطرة في القارئ القرآني، واستخدموه لإنشاء النموذج الصوتي باستخدام مدرب النموذج الصوتي «سفنكس"sphinx train" .

تتوفر أدوات حاسوبية لخدمة القرآن الكريم ويجري التفاعل مع هذه الأدوات باستخدام فأرة الحاسوب أو لوحة المفاتيح، ولكن في بعض الحالات يصبح استخدام تلك الأدوات صعباً، على سبيل المثال عند قيادة السيارة أو في حالة المكفوفين؛ لذلك هدفت هذه الورقة البحثية إلى إنشاء قارئ قرآني يُتحكم به عن طريق الكلام.

قامت هذه الورقة البحثية بوصف عملية تصميم التعرف على الكلام المستمر والموجه

نحو المهام للغة العربية استناداً إلى CMU Sphinx4 وهو عبارة عن مجموعة من أنظمة التعرف على الكلام، وذلك لاستخدامها في واجهة الصوت من القارئ القرآني. Sphinx4 وهو عبارة عن تطبيق يستعمل ماركوف المخفي (HMM) الخاص بالتعرف على الكلام. علاوة على ذلك، فقد عرضت هذه الورقة البحثية الخطوات الأولى نحو تطوير القارئ القرآني الذي يُتحكم به عن طريق الكلام باستخدام الإطار الخاص بنظام التعرف على الكلام sphinx4. في هذه الخطوات، حدد الباحثون الكلمات التي ينبغي التعرف عليها وقاموا بجمع المنظومة المستخدمة لتدريب البرنامج الصوتي باستخدام مدرب النموذج الصوتي .sphinx train

قدمت هذه الورقة مفهوم القارئ القرآني الذي يُتحكم به بواسطة الكلام، وشرحت بالتفصيل طريقة جمع المنظومة والنموذج الصوتي مع مراعاة خصوصيات اللغة العربية والتطبيق المطلوب. استخدم الباحثون مدرب النموذج الصوتي الذي يتطلب إدخال الكلام المسجل، والنسخ، والقاموس والملفات الصوتية لإنتاج النموذج الصوتي. وقد استخدمت العديد من إعدادات مدرب النموذج الصوتي وتكويناته برامج مكتوبة بلغة البرمجة بيرل.

Appl. Math. Inf. Sci. 9, No. 6, 2885-2897 (2015)

Year of publication: 2015

Multimodal Arabic Speech Recognition for Human-Robot Interaction Applications

Alaa Sagheer

التعرف على الكلام العربي متعدد الوسائط للعلاقة بين الإنسان والإنسان الآلي (الروبوت) تطبيقات تفاعلية

المقدمة

نظرا للطلبات المتزايدة على التعاون بين الإنسان و الإنسان الآلي (الروبوت)، من المتوقع أن توفر الروبوتات القدرات الإدراكية على غرار الإنسان. أحد التحديات التي تواجهها الروبوتات هو قدرتها على التواصل مع الناس بعدة أشكال : (١) قدرات السمع (الكلام الصوتي)، (٢) المعلومات البصرية (خطاب بصري) (٣) معرفة هوية المستخدم من خلال التعرف على الوجه. هناك حاجة ماسة لنظام يجمع بين هذه المزايا في إطار واحد ويتغلب على العيوب التقليدية في كل واحدة. سبق وأن قدم الباحث نظاما يجمع بين وحدة التعرف على الخطاب البصري مع وحدة التعرف على الهوية في إطار واحد، أما في هذه الورقة، فقد أضاف الباحث إضافة للتعرف البصري والتعرف على الهوية، أضاف التعرف على الكلام المسموع وبذلك أصبح باستطاعة النظام التعرف على المؤثرات السمعية والبصرية الشاملة . وغني عن القول أن خطاب الصوت هو وسيلة مهمة للاتصال؛ وبالتالي، فإن عدداً كبيراً من التطبيقات يستند إلى الموارد البشرية المعتمدة على الصوت.

في هذا البحث، تستخدم قاعدة البيانات العربية السمعية والبصرية التي وضعها المؤلف، وذلك بزيادة عدد الكلمات لتكون ٣٦ كلمة بدلا من ٩ كلمات فقط. ، يتضمن البحث أيضا

١٣ من العبارات العربية، في حين لم تكن قاعدة البيانات السابقة تتضمن العبارات. بالإضافة إلى ذلك، زيد عدد المواد لتكون في ٥٠ موضوعا بدلا من ٢٠ موضوعا. ومن المؤكد أن زيادة عدد من المواضيع لتصل إلى ٥٠ موضوعا يوسع المعلومات المدرجة في قاعدة البيانات ويضمن تعميم النظام نحو التغيرات في المستخدمين. وفقا لأفضل معرفة متوفرة، قاعدة البيانات هذه هي أول قاعدة بيانات عربية سمعية وبصرية. وقد استعملت ٢٦ من الكلمات العربية المستخدمة في تجارب هذه الورقة، حيث تحتوي كل كلمة بين ٢-٤ فونيمات. تتضمن قائمة الكلمات الأرقام العربية (١،٢،٣،٩)، (١٠،٢٠،٣٠،١٠٠) وأيام الأسبوع (السبت، الأحد، الجمعة) كلها باللغة العربية.

الخلاصة والأعمال المستقبلية

استنادا إلى التعرف على الكلام البصري ، أو قراءة الشفاه، في النظام الذي قدمه المؤلف سابقا، أقترح في هذا البحث ترقية وتعزيزا لهذا النظام، فقد قُدمت خطوات قليلة لتعزيز النظام الموصوف سابقا. أيضا، جرى هنا الجمع بين الإشارات الصوتية والإشارات البصرية من أجل إنتاج نظام عام يعمل على الوقت الحقيقي(real time). وعلى عكس الأنظمة التقليدية حاليا، فإن النظام المقترح يتكيف مع حركات وجه المستخدم التي لا يمكن تجنبها في واقع الحياة. من أجل العمل في الوقت الحقيقي (real time)، استُخدمت خوارزمية البحث السريع لخريطة التنظيم الذاتي (Self Organizing Map (SOM)) من أجل تحقيق مهمة استخراج الميزات، كما تقلل من مساحة الإدخال عالية الأبعاد (high dimensional input space) إلى مساحة ميزة منخفضة الأبعاد. ولمهمة التصنيف، استخدمت خوارزمية أقرب جار(nearest neighbor) ونموذج ماركوف المخفية من أجل التعرف على الميزات المستخرجة. وفي تجارب هذه الورقة، زيد عدد المواضيع ليكون عدد الأشخاص ٥٠ شخصا بدلا من ٢٠ وعدد الكلمات ٢٦ بدلا من ٩ كما كان في النظام السابق. وتؤكد زيادة عدد من الموضوعات والكلمات أن النظام أصبح أكثر شمولا مما يجعله واعدا أكثر مستقبلا.

The 13th International Arab Conference on Information Technology ACIT'2012 Dec.10-13, pp 313-319.

Year of publication: 2012

Dealing with Emphatic Consonants to Improve the Performance of Arabic Speech Recognition

Majed Alsabaan, Iman Alsharhan, Allan Ramsay, Hanady Ahmad

التعامل مع الحروف المطبقة لتحسين أداء التعرف على الكلام العربي

المقدمة

اللغة العربية واحدة من اللغات السامية، ومن الطبيعي أن توجد هناك كثير من الاختلافات إذا ما قورنت مع أي من اللغات الأوروبية. ويرتبط هذا الاختلاف بالنظام الصوتي، فاللغة العربية لديها أربعة حروف تسمى الحروف المطبقة وهي الضاد والطاء والصاد والظاء، ولها ما يقابلها من حروف بسيطة موجودة في اللغات الأخرى، وهي على التعاقب الدال والتاء والسين والذال.

هذا البحث محاولة تطوير نظام تعرُف على الكلمات المقطعة العربية بغض النظر عن المتكلم، بحيث يمكنه التعامل مع الحروف المطبقة الفريدة في اللغة العربية. يستند النظام على نموذج ماركوف المخفي (Hidden Markov Models (HMMs)) باستعمال برنامج خاص لنموذج ماركوف المخفي (HTK). تكونت قاعدة البيانات التي استعملت في التدريب والفحص من ٣٢٠٠ قطعة صوتية (token) بالتعاون مع ٢٠ متكلما باللغة العربية. ولغرض استعمال البيانات بأفضل ما يمكن فقد استُعمل في البحث خمس جولات للتأكد المتبادل لغرض تحسين الصلابة والفحص. بيّنت التجارب أن أداء النظام كان ٦ , ٢٢٪ عند استعمال الكلمات كوحدات صوتية في نمذجة الصوت، وهذا يبين أن الحروف المطبقة مصدر رئيس للصعوبة في الوصول إلى نظام تعرف على الكلام العربي. كما أن البحث توصل إلى أنه باستعمال الفونيمات كنماذج صوتية،

فإن الدقة يمكن أن ترتفع إلى ٤ , ٦١٪. يقدم البحث طريقة فريدة لتقليل سوء الفهم باستعمال توصيف صوتي بمواصفات متدنية (underspecified phonetic transcription) للكلام المدخل. وبذلك أصبحت نسبة الأخطاء الكلية ٢٣٪. لقد أثبت البحث أن هذا الأسلوب يمكن استعماله في نظم تعليم اللغة وبالأخص لغير الناطقين بالعربية لغرض إنتاج أصوات صحيحة لتعليم اللغة.

التدريب والاختبار

اختيرت مجموعة من الكلمات العربية الحقيقية التي يحتوي نصفها على حروف مطبقة، ويحتوي النصف الآخر على نظراء تلك الحروف. تحتوي البيانات على مجموعة من الكلمات، حيث تأتي الحروف المطبقة ونظراؤها في السياق نفسه، وتحتوي أيضا على مجموعة من الكلمات بحيث تقع تلك الحروف (في البداية، والوسط، والنهاية). وقد طُلب من المتكلمين تكرار كل كلمة ٥ مرات، وبذلك بلغ العدد الإجمالي للكلمات ٣٢٠٠ كلمة، جُمعت من ٢٠ ناطقا عربيا من جنسيات مختلفة: ثلاثة من الكويت وثلاثة من مصر، وثلاثة من سوريا، واثنان من السعودية، واثنان من اليمن، واثنان من الأردن، واثنان من السودان، وواحد من البحرين، وواحد من العراق، وواحد من فلسطين. وكان اثنا عشر من هؤلاء المتحدثين من الذكور، وثمانية من الإناث. وجميعهم تتراوح أعمارهم بين ٢٥ و ٣٥ عاما.

ولضمان حصولنا على نتائج موثوقة، استخدمنا نهجا مزدوجا للتحقق من صحة الاختبار كوسيلة لتقييم النظام المقترح، وذلك بالتقسيم العشوائي للبيانات إلى ٥ مجموعات فرعية متساوية الحجم وتقسيم الأداء على ٤ مجموعات فرعية، والتحقق من صحة الأداء على المجموعات الأخرى. وتكررت هذه العملية ٥ مرات مع كل من المجموعات الفرعية الخمسة التي استعملت مرة واحدة فقط كبيانات للتحقق. وفي النهاية أُخذ معدل المجموعات الخمسة. وتتمثل ميزة هذه الطريقة على نهج التقييم المعياري في أن جميع الرصدات (observations) تستخدم لكل من التدريب والاختبار، مما يعطي اختبارا أكثر قوة للتجارب مع مجموعات البيانات الصغيرة. وقد استُعمل برنامج HTK لتحليل النتائج.

النظام المجهز في هذا البحث يتعرف على الأصوات العربية الخاصة بالحروف المطبقة باستعمالHTK. ويحوي مرحلتين رئيسيتين: الأولى مرحلة التدريب، وذلك ببدء وتخمين أكبر احتمالية بعوامل نموذج HMM، وهي تحتاج إلى بيانات صوتية مرفق معها مقاطع الكلام المصاحبة لها.

المرحلة الثانية هي مرحلة الفحص لكي تجري مقابلة الكلام المدخل مع شبكة HMM واسترجاع نسخة لكل من إشارات الصوت.

في هذا البحث قمنا بخمس تجارب رئيسية؛ التجربتان الأوليان تبينان كيفية تأثير استعمال الوحدات الصوتية على دقة التعرف. في التجربتين الثالثة والرابعة قمنا بتدريب النظام وفحصه على ذكور فقط، وإناث فقط، بنفس حجم البيانات، والهدف من ذلك هو معرفة تأثير تغير الجنس على دقة التعرف. من هذه التجارب قمنا بتحليل مصفوفة الوهم (confusion matrix) لمعرفة أي من الفونيمات يقع فيها الوهم أكثر من غيرها، ومحاولة البحث عن حلول لتقليل التوهم.

لقد بينت الدراسة أن حرف الضاد هو أكثر الحروف المطبقة التي يصعب نطقها وتُسبب توهما، وهذا التوهم والتشويش مع صعوبة النطق بالنسبة للناطقين باللغة العربية أنفسهم.

كما أن البحث توصل إلى نتيجة تشير إلى اختلاف التمييز بين الذكور والإناث، حيث وُجد أن أداء الإناث كان أفضل من أداء الذكور.

يقترح البحث زيادة كمية البيانات الخاصة بالتدريب التي تحتوي على هذه الفونيمات التي تسبب الارتباك لتأكيد هذه النتائج بشكل أوضح.

النتائج

تستند النتائج الواردة هنا إلى نتائج نظام التعرف الآلي على الكلام التلقائي الموصوف أعلاه. وقد عُرضت نتائج التعرف التلقائي على الكلام في أربعة أقسام فرعية. القسم الفرعي الأول يعطي نتيجة للتجارب على مستوى كلمة وصوت، ويناقش القسم الفرعي الثاني نتائج

التجارب المعتمدة على نوع الجنس، ويستعرض القسم الفرعي الثالث الصوتيات المحيرة التي لوحظت من خلال هذه التجارب.

المقترحات المستقبلية

الغرض من الدراسة الحالية هو التحقق من العوامل التي تؤثر على نتائج تدريب متلقي الكلام لتحديد الحروف المطبقة العربية بشكل صحيح. ومع ذلك، كان هذا التحقيق محدودا بسبب حجم البيانات التدريبية الصغيرة.

ولذلك، يلزم إجراء المزيد من التحقيقات التجريبية. إضافة إلى ذلك، فإن الأساليب المستخدمة لإظهار الحروف المطبقة يمكن أن تُطبق على الصوتيات الأخرى مثل حروف العلة، وحروف البلعوم، والحروف الساكنة التي قد تكون أيضا عقبة لتطوير نظام التعرف على الكلام. ويمكن أن يستفاد من نتائج هذه الدراسة في عدد من التطبيقات المهمة المستقبلية ، على سبيل المثال مساعدة متعلمي اللغة العربية من غير الناطقين بها لإنتاج الأصوات بشكل صحيح.

٣-٤-٢ أبحاث القارئ الآلي

وتضم ٥ أبحاث بينها بحث مسحي بعنوان: تمييز الحَرف العَربي - دراسة مسحيّة.

وبحث من فئة (أ) عنوانه : التعرف على الكلمات العربية المكتوبة بخط اليد باستعمال التقطيع إلى حروف والشبكات العصبية المتكررة.

وثلاثة أبحاث فئة (ب)، هي: التعرف على الحروف العربية باستعمال الخطوات المتعاقبة التقريبية، واختيار الميزات المثلى للتعرف على الحروف العربية المكتوبة بخط اليد، وتقنية التعرف الضوئي على الحروف بعد معالجة تصحيح الأخطاء باستخدام الاقتراحات الإملائية لجوجل.

International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 8, No. 2 (2015), pp. 401-426, http://dx.doi.org/10.14257/ijsip.2015.8.2.37

Year of publication: 2015

A Survey on Arabic Character Recognition

Ahmed Lawgali

تمييز الحَرف العَربي – دِراسة مَسحيّة

المقدمة

تُستخدم تقنيات التعرّف التّلقائي (الأوتوماتيكي) والتّمييز بين الحروف والكلمات بواسطة الحاسب الآلي لقراءة محتوى الوثيقة المطبوعة أو المكتوبة (المخطوطة) على نطاق واسع وضمن تطبيقات مختلفة، منها: الفرز الآلي للبريد العادي، أو معالجة الشّيكات البنكية أو تحرير الوثائق القديمة. وتعتمد هذه التقنية على التعرّف البصري على الحروف (Character Optical Recognition) الذي يُعرّف بأنه: «تحويل صورة النص المكتوب إلى نص قابل للتعديل لتجنب إعادة كتابته."

ويمكن تقسيم تقنية التعرّف على الحروف أو تمييزها إلى نوعين: تقنية أثناء الكتابة (On line OCR) ، وتقنية التعرف على النصوص المكتوبة (Off line OCR) ، ففي النّوع الأول يتمّ التعرف على الحروف أثناء عملية الكتابة باستخدام تقنية التتبع الرقمي للقلم، أما في النوع الثاني فيجري التعامل مع الصّور الممسوحة ضوئياً للوثائق المكتوبة مسبقاً. ومن الجدير بالذّكر أن هناك العديد من اللغات التي تستخدم الحروف العربية مثل اللغات الأوردية والفارسية والجاويّة التي يمكن أن يخدمها التعرف على الحروف العربية بالإضافة طبعا إلى اللغة العربية.

تضمنت هذه الدراسة أربعة مباحثَ رئيسيّة: أولاً: التحديات والدوافع للتعرّف على

الحروف العربية، ثانياً: مقدّمة عن نظام التعرف على النص العربي المكتوب بخط اليد بنوعيه، ثالثاً: مراجعة أدبيات أنظمة التعرف على الحروف العربية، التي تصنف إلى فئتي أنظمة التعرف على الحروف: المطبوعة والمكتوبة بخط اليد، رابعاً: ناقشت الدراسة اتجاهات العمل المستقبلية في هذا المجال، وأخيراً الخاتمة.

مُشكلة البحث والمُبررات:

أثبتت التّجربة عدم صلاحية الأنظمة الحالية في التعرف أو تمييز الحرف العربي لمعالجة النّص العربي الكتابي (المخطوط) وما ينطوي عليه من تعقيدات تفوق قُدرة هذه الأنظمة. وأظهرت البحوث التي أُجريت على هذه المشكلة أن معظم الأساليب أو التقنيات التي جرى تطويرها صُمّمت على أساس التعرّف على كامل الكلمة أو المفردة العربية دون تجزئة أو تقطيع، بسبب صعوبة ذلك والطبيعة الخاصة لترابط حروف الخط العربي وتشابكها.

وقد حاول بعض الباحثين التعرف على الحروف العربية بعد تجزئة أو تقطيع الكلمات إلى حروف، إلاّ أن التعرّف على الحروف أو تحديدها لم يحُلّ المشكلة لأنه قد يُعطي أحياناً – تصنيفاً خاطئاً لبعض حُروف الكلمة. غير أننا، بالمقارنة مع الأنظمة أو الأساليب الخاصة بالتعرف على الحروف المكتوبة بخط اليد للغتين اللاتينية والصينية، نجد القليل من البحوث حول استخدامات مثل هذه الأنظمة في التعرف على الحروف المكتوبة بخط اليد باللغة العربية.

المباحث الرئيسية للدراسة

المبحث الأول: التحديات

يمتاز النص العربي بأنه نص متشابك؛ أي أن حروفه متصلة ببعضها من خلال خط (وهمي) يُعرف بخط الأساس (أو السطر) (baseline) وهناك أيضاً خُطوط أخرى تظهر فوق خط الأساس (السطر) وتحته تعرف بـ الصّاعد (ascender) والهابط(descenders) ، وهناك ستة أحرف هي (و، ز ، ر، ذ، د، ا) ليس لها شكل مُحدد في بداية الكلمة ومنتصفها، وبالتالي

لا تتصل هذه الحروف بالحروف التي تليها في الكلمة، وهو ما يُجزّئ الكلمة الى مقاطع (sub words). وهناك حروف ذات شكل متشابه لكنها تتمايز باختلاف عدد النقاط الظاهرة عليها وموقع تلك النقاط سواء في الأعلى أو في الأسفل، مثل (ث، ت، ب)، كذلك الهمزة يمكن استخدامها بعدة طرق أو مواضع بحيث تُعطي خمسة حروف مُختلفة أو يمكن استخدامها بشكل مُنفرد.

من جهة أخرى، يمكن كتابة الكلمة نفسها – باللغة العربية – بعدة أساليب ووفقاً لأحجام مُختلفة. كذلك هناك بعض الحروف في اللُغة العربية المكتوبة بخط اليد يمكن أن تتّحد عمودياً لتشكّل ما يُعرف بـأداة الرّبط أو الوصل، وهو ما يعني أن الحرف الثاني قد يظهر قبل الحرف الأول في بعض الحالات. ومن الأمثلة على ذلك حروف (ح، خ، ج، م، ل) التي قد تظهر بعد بعض الحروف الأخرى. وفي حالات أخرى قد يظهر التداخل بين بعض الحُروف العربية في الكلمة مثل حروف (ع، ا، ر) دون أن تُلامس بعضها، كما هو الحال في كلمة (الذراع).

وهناك بعض الحروف التي تظهر متشابهة في النص العربي المكتوب بخط اليد، في حين أنها مختلفة، ورغم اختلافها يصعب تمييز اختلافها حتى باستخدام العين المجرّدة. وهناك اختلاف بين الطول والعرض في كتابة الحرف العربي كما هو الحال في حرفي (ا، ب).

المبحث الثاني: نظام التعرف على الحرف العربي

النوع الأول من أنظمة التعرف على الكتابة العربية أثناء الكتابة الفعلية بخط اليد، هو ما يحتاج إلى الكتابة على لوح خاص باستعمال قلم خاص . وبالطبع لا يُمكن الاعتماد على هذا النظام في التعرف على الوثائق المكتوبة مسبقاً. أما نظام التعرف على الحرف العربي المكتوب بخط اليد مسبقا فهو يتعامل مع الصُور الممسوحة ضوئياً.

ويمكن تصنيف هذه الأنظمة إلى فئتين: أنظمة التعرف على الحروف المطبوعة، وأنظمة التعرف على الحروف المكتوبة بخط اليد، وهذه الأخيرة يمكن أيضاً تصنيفُها إلى فئتين: أنظمة التعرف المبنية على التجزئة أو التقطيع (segmentation) وأنظمة التعرف غير المبنية على التجزئة أو التقطيع.

وينطوي نظام التعرف على الحرف العربي المكتوب بخط اليد على العديد من التّحديات التي تُعزى إلى التعقيد والغموض في أساليب الكتابة.

وفي حين أن الأحرف المطبوعة لها نمط واحد وحجم مُحدد لكل نوع من الخطوط (fonts)، تختلف طبيعة الأحرف المكتوبة بخط اليد في أنّ لها أنماطاً وأحجاماً مختلفة، سواء للكاتب نفسه أو لعدّة كُتاب. ويمكن التعرف على الكلمات المكتوبة بخط اليد بطريقتين: إمّا من خلال التعرف على الكلمة كاملةً دون تجزئة، أو من خلال التعرف على الكلمات المكتوبة على أساس التجزئة (التقطيع). ولما كانت عملية التجزئة (تقطيع الكلمة) هي المصدر الرئيسي للأخطاء في أنظمة التعرف، عمدت معظم هذه الأنظمة إلى تجنّب هذه الخطوة والاعتماد فقط على التعرّف على كامل الكلمة المكتوبة دون تجزئة أو تقطيع.

من جهة أخرى، ناقشت الدراسة في هذا المبحث خطوات التعرف على النص المكتوب بخط اليد، وهي:

المُعالجة المُسبقة (preprocessing) للنص: الغرض من هذه الخُطوة – في نظام التعرف على النص المكتوب بخط اليد – هو تعزيز قابلية صورة النص للقراءة وإزالة التفاصيل التي ليس لها قوة تمييزية يُمكن الإفادة منها في عملية التعرّف على الحروف العربية المكتوبة. وتتضمن هذه الخطوة – عادةً – عدة مهام، هي : استخدام الثنائية (الزوجية) (binarisation)، وإزالة الضوضاء (noise removal) ، وتتبّع خط الأساس (baseline detection)، والتطبيع (normalization). وفيما يلي شرح موجز لكل منها:

استخدام الثنائية (الزوجية)، وإزالة التّداخل (الضوضاء): وهي عملية تستخدم لتحويل الصورة النصية إلى تنسيق ثنائي؛ من خلال إعطاء قيم بكسل الخلفية الرقم (١) التي تمثل اللون (الأبيض) وإعطاء قيم بكسل الأمامية الرقم (صفر) الذي يمثل اللون (الأسود). وتُنفّذ هذه العملية عن طريق اختيار قيمة كفؤة لطريقة العتبة (thresholding method)، وإحدى مزايا هذه العملية هي أنها تزيد من سُرعة المعالجة. يحدث التشويه في الصورة عادة أثناء عملية المسح الضوئي. ويمكن اعتبار الأجسام الصغيرة التي لا تشكل جُزءاً من الكتابة مصدراً للضوضاء

ويجب إزالتها. هناك بعض الأساليب المستخدمة للحد من الضوضاء، وتشمل هذه الأساليب التّصفية (التّرشيح)(filtering) والعمليات الصّرفية. ويمكن تصميم العديد من المرشحات الموضعية (المكانية) والتكرارية لأغراض مختلفة، مثل التنعيم (smoothing) وإزالة الضوضاء ((removing noise ، حيث يقلل التنعيم من الضوضاء باستخدام عمليات التشكل الرياضية (operations mathematical morphology). وعادة يتم تحقيق ذلك بطريقتين: إما عن طريق فتح المساحات بين الأشياء المتّصلة أو عن طريق سد الثغرات الصغيرة. وكلا الطّريقتين سواء الفتح والإغلاق، يمكن تطبيقها على العمليات الصّرفية.

تتبّع (كشف) خط الأساس: ذكرنا سابقاً أن خط الأساس هو الخط الوهمي الذي يربط حروف الكلمة ببعضها، وهذه الخطوة من أهم خطوات مرحلة المعالجة. ويُساعد الكشف عن خط الأساس في تحديد بعض السمات التركيبية، مثل النقاط ومواضعها، والصّعود والهُبوط، فضلاً عن تصحيح الالتواء / الميل (correction skew/slant)، وهناك عدة طرق وتطبيقات مُختلفة للكشف عن خط الأساس.

التّطبيع (Normalisation): تمتاز الكتابة اليدوية بأنّ لها أنماطا وأحجاما مُختلفة. لذلك، تُعتبر عملية التّطبيع واحدة من أهم المهام في عملية التعرف على النص، وتُستخدم للحدّ من الاختلاف بين الصُور النّصية وضبط حجم الحرف أو الكلمة. ومن الشائع استخدام تطبيع حجم النص لتقليل حجم الاختلاف، وضبط حجم الأحرف أو الكلمات لتصبح ذات حجم متماثل.

تجزئة (تقطيع) النّص

تقتضي مرحلة التجزئة تقسيم النص إلى وحداته الفرعية، مثل الخُطوط أو الكلمات أو الأحرف. ويُعدُّ التقسيم مرحلة مُهمّة لأن له تأثيرا على مُعدّل أو مستوى التعرّف. وهناك عدة أنواع من التّجزئة:

- تجزئة الصّفحة إلى خُطوط أو سُطور.

- تجزئة السّطر إلى كلمات.
- تجزئة الكلمات إلى حروف.

كذلك هناك عدة تقنيات مُطبّقة لتجزئة الحروف:

تقنية التجزئة على أساس الإسقاط الرأسي (Vertical Projection)؛ وهذه التقنية تعتمد على اختزال المعلومات ثنائية الأبعاد إلى بعد واحد.

تقنية التجزئة المعتمدة على التنحيف (Thinning)؛ من خلال استخدام بعض الخوارزميات التي تُجرّد الحرف للحصول على هيكله الأصلي (skeleton).

تقنية التّجزئة المعتمدة على التتبّع أو الاقتفاء الكونتوري (المعالم) (Contourracing): وتتحقق هذه التقنية من خلال تتبع المحيط الكنتوري (الخارجي) (outer contour) للجسم الرئيسي للكلمة.

تقنية التجزئة بناء على الشبكات العصبية الاصطناعية (Artificial NeuralNetworks): وتُستخدم هذه التقنية للتحقق من نقاط التجزئة الصالحة.

التقنية المعتمدة على العوامل الصرفية (Morphological Operations) .

تقنية استخراج الخاصية (Feature Extraction).

تقنية التصنيف: تحاول هذه التقنية القيام بمهمة التعرف على الكائن أو تحديده، من خلال مقارنة صفاته (خواصه) بقائمة أو بمجموعة أشكال أخرى. وتفترض هذه التقنية وجود اعتياد أو تمرّن على التعامل مع ذلك الكائن المراد التعرف عليه أو تحديد صفاته. لذلك، يستخدم المصنّف (classifier) لتحديد الكائن باستخدام ميزاته، ثم يتم مقارنتها وحفظها كنماذج للأشكال المعتادة (المتمرّن عليها). ويتم استخراج ملامح الكائن في مرحلة الاختبار ومقارنتها مع ملامح النماذج المتمرّن عليها للتحديد أو التعرف على الكائن المجهول. وتعتمد الطرق الشائعة للتصنيف على ما يلي:

الشبكات العصبية الاصطناعية.

نموذج ماركوف الخفيّ.

خوارزمية الجار الأقرب (K-nearest neighbor).

المبحث الثالث: مُراجعة أدبيّات أنظمة التعرف على الحروف العربية

استعرض هذا المبحث الأعمال السابقة عن أنظمة التعرف على الحروف العربية، حيث بيّنت الدراسة أن هناك القليل من الأعمال البحثية في مجال التعرف على الحروف العربية المكتوبة بخط اليد، بالمقارنة مع نظيراتها اللاتينية والصينية. ويظهر استعراض الأدبيات ذات الصلة بأنه يمكن تصنيف أنظمة التعرف على الحروف العربية إلى فئات مختلفة، ويمكن التعرف على النصوص العربية المطبوعة والنص العربي المكتوب بخط اليد، إلا أنه لتطوير هذه الأنظمة، يلزم وجود قاعدة بيانات للأحرف لتسهيل العملية. ونظراً لعدم وجود قاعدة بيانات مرجعيّة شاملة ومتاحة للجمهور، فقد طوّر مُعظم الباحثين نُظمَهم الخاصة استناداً إلى مجموعة من البيانات التي جمعوها بأنفسهم.

المبحث الرابع: اتجاهات العمل المستقبلية (التوصيات)

جرى التعامل في هذه الدراسة مع مشكلة التعرف على الكلمات العربية المكتوبة بخط اليد باستخدام الأنظمة التي تستعمل النص المكتوب. وبشكل عام، هناك أسلوبان مُستخدمان لتقنية التعرف على خط اليد في هذه الحالة، الأول: هو تصنيف الكلمة بأكملها دون تجزئة؛ والآخر هو التصنيف على أساس تقسيم الكلمة إلى أحرف ثم تصنيف الحُروف تباعاً. ونظراً للطّبيعة المتشابكة أو المتداخلة للنص العربي المكتوب بخط اليد، فإن ذلك يجعل تقسيم كلماته إلى حروف فردية مُهمة صعبة. لذلك لم تُفلح معظم خوارزميات التجزئة المقترحة حالياً في حلّ مُشكلة تداخل الأحرف في الكتابة اليدوية العربية، كما أنها لا تستطيع التعرّف على الكلمة بعد تجزئتها. ولا تزال تقنيات التجزئة تمثل تحدياً في التعرف على النُصوص العربية وتحتاج إلى تحسين. ومن ثم، فقد طرح الباحثون عدة طُرق على أساس التعرّف على كامل الكلمة دون

تجزئة، فضلاً عن غيرها من الأساليب التي تفترض بأن الأحرف قد جُزّئت أو قُطّعت بالفعل، وذلك لتجنب عملية التجزئة. غير أنّ استخدام هذا الأسلوب - أي التعرف على الكلمات الكاملة دون تجزئة - له مُحدّدات، فيُمكن أن يصنف فقط الكلمات المعتادة (المتمرّن عليها)؛ في حين أن أسلوب التعرّف القائم على التجزئة لا يتضمن ذلك المُحدّد.

الخاتمة:

تناولت هذه الدراسة المسحيّة أنظمة التعرُف على الحرف العربي، حيث استعرضت الدراسة أهم الأساليب المستخدمة في تمييز النص العربي المطبوع، التي أثبتت التجربة عدم صلاحيتها لمعالجة النّص العربي المكتوب بخط اليد وما ينطوي عليه من تعقيدات تفوق قدرة هذه الأنظمة. وأظهرت البحوث التي أجريت على هذه المشكلة أن معظم الأساليب أو التقنيات التي جرى تطويرها صُمّمت على أساس التعرف على كامل الكلمة أو المفردة العربية دون تجزئة أو تقطيع، وذلك بسبب الطبيعة الخاصة لتداخل وتشابك حروف الخط العربي. وصنفت الدراسة أنظمة تمييز الحرف العربي الى قسمين: المطبوع والمكتوب بخط اليد. واختبرت الدراسة أهم الأدبيات والأعمال السابقة على تمييز حروف النص العربي دون تجزئة أو تقطيع للكلمة، وناقشت الدراسة الخوارزميات المستخدمة في تقسيم الكلمات إلى حُروف. وتعتبر مرحلة التجزئة هي المرحلة الأكثر صعوبة والمصدر الرئيسي للأخطاء في أنظمة التعرف على النص العربي.

بالإضافة إلى ذلك، معظم خوارزميات التجزئة المقترحة حالياً لم تستطع حل مشكلة تداخل الحروف في الكتابة اليدوية العربية، كما أنها لم تنجح في التعرف على الكلمات العربية بعد تجزئتها بشكل كبير.

International Journal on Document Analysis and Recognition, Volume 17 Issue 3, September 2014, Pages 275-29, DOI=http://dx.doi.org/10.1007/s10032-014-0218-7

Year of publication: 2014

Recognizing handwritten Arabic words using grapheme segmentation and recurrent neural networks

Gheith A. Abandah, Fuad T. Jamour, Esam A. Qaralleh

التعرف على الكلمات العربية المكتوبة بخط اليد باستعمال التقطيع إلى حروف والشبكات العصبية المتكررة

لقد اعتنى بحث سابق للمؤلفين بالتعرف على الحروف المنفصلة فقط. أما عند استعمال النظام للتعرف على الكلمات فيجب إضافة مرحلة تقطيع الكلمة إلى حروف كمرحلة تمهيدية. وفيما يلي بعض التفاصيل الواردة في البحث.

يحتوي النظام على خمس مراحل:

المرحلة الأولى تجزئة الكلمة إلى أجزائها (**sub-word segmentation**).

المرحلة الثانية تقطيع جزء الكلمة إلى حروف (**grapheme segmentation**).

المرحلة الثالثة استخلاص الميزات (**extraction feature**).

المرحلة الرابعة ترجمة متسلسلات الميزات إلى متسلسلات حروف الكلمات (**sequence transcription**) وهذه المرحلة لا تستعمل نموذج اللغة (**language model**).

المرحلة الخامسة مواءمة الكلمات باستعمال معجم كلمات (**word matching**) وهي مرحلة اختيارية ويمكن إضافة كلمات أو حذفها من المعجم لزيادة نسبة التعرف على الكلمات بدون الحاجة لإعادة تدريب الشبكة العصبية المستخدمة في ترجمة المتسلسلات.

تجزئة الكلمة

تجري هذه المرحلة بخطوتين: الخطوة الأولى بتجزئة الكلمة إلى أجزائها المكوّنة من مقاطع تتكون من حرف أو أكثر متصل مع حروف أخرى، ثم تجزئة الأخيرة إلى حروف (graphemes). في بداية الخطوة الأولى يُحدد الخط الأساسي (baseline) ثم يجري تحديد الأجزاء الأساسية من أجزاء الكلمات (جسم الجزء) من الأجزاء الثانوية (النقاط والحركات).

يتحدد الخط الأساسي باستعمال طريقة منحنى تسقيط أفقي (horizontal projection histogram method) وارتباط الجزء الثانوي مع الجزء الأساسي. فيكون الخط الأساسي عند قمة منحنى التسقيط الأفقي. كما استعملت خوارزمية لاستخلاص الأجزاء المتصلة بالاعتماد على الخط المحيط (contour) وذلك لتحديد جسم الحرف وتوابعه من النقاط. وتحدد النقاط بأنها أجسام صغيرة منفصلة بعيدة عن الخط الأساسي. وتقع فوق الجزء الأساسي للكلمة أو تحته.

تجزئة الحروف

تضم هذه العملية خطوتين هما الكشف عن النقاط المميزة (detection Feature point)، ويجري التعرف على ميزات الحروف بواسطة الخط المحيط لجزء الكلمة. تهمل الأجزاء الثانوية مؤقتًا قبل استعمال طريقة التنحيف المعروفة باسم خوارزمية دويج للتنحيف (Deutsch's thinning algorithm) وهي خوارزمية جيدة لكنها قد تفقد الحافات الدقيقة مثل أسنان حرف السين. إن النقاط المميزة التي تُكشف هي نقاط النهاية ونقاط التفرع ونقاط التقاطع باستعمال طريقة التجاور الثمانية لكل نقطة من نقاط الهيكل (skeleton) الناتج من عملية التنحيف. ثم بعد ذلك تربط ميزات الحرف مع الجزء الثانوي تمهيدا للمرحلة اللاحقة.

استخلاص الميزات

تستخلص الميزات الإضافية التي يبلغ عددها ١٠٣ ميزة، ويمكن تجميعها إلى ست

مجموعات، هي المزايا الإحصائية (Statistical features)، ومزايا الوضعية (Configuration features)، ومزايا الهيكل (skeleton)، ومزايا المحيط (Boundary features)، وواصفات فورير البيضوية descriptor elliptic Fourier))، ومزايا الإتجاه (Directional features).

عملية ترجمة المتسلسلات (Sequence transcription)

تجرى هذه العملية بإيجاد توزيع الاحتمالية (probability distribution) لكل إمكانيات التسلسل. ويجري ترتيب الاحتمالية الأكثر توقعا واختيارها لتمرر إلى المرحلة اللاحقة.

الذاكرة ذات الاتجاهين للعبارات القصيرة الطويلة (Bi-directional long short-term memory)

استعملت طريقة الشبكات العصبية المتكررة (Recurrent neural networks) ذات الاتجاهين (bidirectional) لكي تتمكن الطبقة الخارجية من استعمال المحتويات السابقة واللاحقة لكل نقطة ضمن التسلسل.

طبقات أخذ نماذج فرعية (layers Subsampling)

باستعمال الطبقات المخفية المتعددة ذات الاتجاهين (bidirectional RNN hidden layers) يمكن تحسين الأداء بالمقارنة مع المعمارية ذات الطبقة الواحدة، لكن ذلك يمكن أن يؤدي إلى تضخم في حجم الارتباطات بين الطبقتين الأمامية والخلفية. ولغرض تقليل هذا التضخم تفصل الطبقات الأمامية إلى مستويات فرعية.

تُفصل الطبقتان في كل مستوى تغذية أمامية إلى طبقات جزئية تغذى بعد ذلك من الأمام إلى طبقتين في المستوى الأعلى. لذلك فإن مجموع أعداد أوزان المستويات المتوسطة يمكن السيطرة عليها بمعايرة حجم المستويات الجزئية.

التبويب المؤقت للارتباطات (Connectionist temporal classification)

يجري تدريب الشبكات للحصول على حالة أفضل احتمالية توزيع للمواصفات الخارجة بالنسبة للمدخلات المتسلسلة كاملة. وهي لا تحتاج إلى تدريب البيانات المجزأة مسبقا.

قياسات الخطأ

يجري فحص مجموعة من نسب أخطاء التسمية (label error rate)، وهي مجموع الإضافات والمحذوفات والاستبدالات في مجموعة الفحص مقسومة على المجموع الكلي للتسميات في مجموعة الفحص مضروبة بـ ١٠٠. وبالطبع فإن الطريقة الفضلى هي احتساب نسب أخطاء التسمية للكلمة أو للسلسلة.

مواءمة الكلمات (Word matching)

عند توفر معجم هناك حاجة إلى خوارزمية لمواءمة الكلمات التي تحسن الدقة تحسناً كبيراً. الخوارزميتان اللتان استعملتا تختاران أفضل عدد من الكلمات في المعجم باستعمال نتائج الشبكات العصبية المتكررة للمراحل المتعاقبة.

مسافة التحرير الموزونة (Weighted edit distance)

هي خوارزمية تعتمد على تحرير أو تعديلها المسافة بين تسلسلين. فالمسافة القياسية المحررة تعرف بأنها أقل عدد كلي من التغيرات والإقحامات والحذوفات المطلوبة لنقل نموذج من وضع إلى وضع آخر.

مواءمة كلمات الشبكات العصبية المتكررة

هناك خوارزمية لهذه المواءمة تقوم بعملية تكامل للطبقات الخارجة.

النتائج التجريبية

استعملت قواعد بيانات خاصة للكتابة بخط اليد، التي كانت قد استخدمت في أكثر من ١١٠ أبحاث منشورة مسبقا في ٣٥ بلدا، وتحتوي على ٣٢٤٩٢ كلمة عربية بخط اليد مكتوبة من أكثر من ١٠٠٠ شخص. هذه الكلمات هي ٩٣٧ اسمًا لمدن وقرى تونسية. وقد استعملنا خمس مجموعات بيانات للتدريب والتحقق والفحص.

التحقق من خوارزمية تجزئة الكلمات

لقد تأكّدنا من صحة التجزئة بفحصها بالنظر، وقد نشر في البحث بعض الأشكال التي تدعم ذلك. كان عدد الكلمات ١٠٧ كلمات ووجد أن التجزئة كانت صحيحة في ١٠٣ كلمات، وأن عدد الكلمات التي جزئت أكثر من الحاجة بلغ ٣ كلمات، والتي كانت أقل من الحاجة كلمة واحدة.

لقد احتوت الشبكات العصبية المحتوية على ثلاثة مستويات أعداد عُقدها ١٠٠ و ١٠٠ و ٣٦٠ مع مستويين جزئيتين، حجماهما ١٢٠ و ١٨٠ عقدة. وقد أنتجت هذه البيانات خطأ مقداره ١٩,٨٪. وعندما دربت على ١٦٩٦ كلمة وصلت الدقة إلى ٩٦, ٩٨٪ .

عرض البحث مقارنة دقيقة مع عدد من فحوصات في أبحاث أخرى مع تبيان مزايا الطريقة المشروحة في البحث.

Arabic Language Resources and Evaluation – Status and Prospects, LREC2002, 1st June 2002. Las Palmas de Gran Canaria.

Year of publication: 2002

Arabic Character Recognition using Approximate Stroke Sequence

Mohammed Zeki Khedher, Gheith Abandah

التعرف على الحروف العربية باستعمال الخطوات المتعاقبة التقريبية

استنادا إلى الإحصائيات السابقة، خلص البحث إلى أن أسلوب التعرف على الحروف العربية يمكن أن يتميز بالمزايا الآتية:

- المقاطع المكونة من حرف واحد لا تحتاج إلى تقطيع.
- المقاطع المحتوية على حرفين تحتاج إلى تقطيع إلى حرفين، ويكون الأول بشكل الحرف الأول والثاني هو الأخير.
- المقاطع المحتوية على ثلاثة أحرف فأكثر تُقطّع إلى الحرف الأول بشكل الحرف الأول، والأخير بشكل الحرف الأخير، وما تبقى بينهما هو حرف وسطي أو أكثر من حرف وسطي.

استند البحث إلى نص طلبت كتابته من ٤٨ شخصا مختلفا ثم أُجري عليه تقطيع يدوي لفرز الحروف المتشابهة المكتوبة من الأشخاص كافة بكل أشكالها المنفصلة والأولية والوسطية والآخرية. ثم بعد ذلك طبقت عليه خوارزمية مطابقة السلسلة بخطوات متعاقبة تقريبية (String Matching Approximate Stroke Sequence) باستعمال الخطوات ذات الثمانية اتجاهات (8-direction stroke Convention) وقد جُرّب على الحروف المنفصلة، وحصلت النتيجة على دقة مقدارها ٨٠٪ لحروف الألف والصاد والهاء، وكانت النتيجة أقل من ذلك لحروف أخرى.

World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering, Vol:1, No:4, 2007

Year of publication: 2007

Optimizing Feature Selection for Recognizing Handwritten Arabic Characters

Mohammed Z. Khedher, Gheith A. Abandah, and Ahmed M. Al-Khawaldeh

اختيار الميزات المثلى للتعرف على الحروف العربية المكتوبة بخط اليد

يختلف نظام الكتابة باللغة العربية عن الأنظمة باللغات الأخرى، حيث يتكوّن النظام العربي من (٢٨) حرفاً تكتب من اليمين إلى اليسار، وأغلب هذه الأحرف تتصل بالحرف الذي يليها والحرف الذي يسبقها باستثناء ستة أحرف يمكن أن تكون مرتبطة مع الحرف الذي يسبقها فقط. يهدف هذا البحث إلى تقديم طريقة للتعرف على الكلمات العربية المكتوبة بخط اليد، وبما أن أنظمة التعرف على الكلام المكتوب تعتمد بشكل أساسي على الميزات (features) المستخرجة من الأحرف؛ فقد ركزت هذه الطريقة على بناء نظام للتعرف على الحروف العربية المكتوبة بخط اليد عن طريق الميزات الخاصة بكل حرف غير متصل (off-line system).

اشتملت الطريقة المقترحة على ثلاث خطوات رئيسة هي المعالجة المسبقة (preprocessing) واستخراج الميزات (feature extraction)، والتعرف (recognition). وفيما يلي توضيح كل منها:

خطوة المعالجة الأولية: تهدف إلى تحضير الصورة المحتوية على الحروف للمراحل التالية عن طريق إزالة الزوائد والمكونات غير المهمة والعيوب التي تحيط بالأحرف ولا تؤثر على هوية الحرف، وهذه الخطوة تشتمل على أربع مهمات أساسية: أولها إزالة الشبكة (Grid removal) ، وهي مهمة تعنى بإزالة الخطوط الأفقية والعمودية، تليها مهمة إزالة التشوهات أو الضوضاء (Noise removal) التي تقوم بتحويل الصورة الملونة إلى ثنائية (أبيض وأسود)، ثم مهمة

التقسيم (Splitting) وفيها تُفصل الأحرف إلى صور مستقلة، وفي المهمة الأخيرة يُرقق سمك خط الحرف إلى بكسل واحد وتسمى هذه المهمة الترقيق أو التنحيف (Thinning).

الخطوة الثانية تُستخلص فيها الميزات والخصائص التي تميز كل حرف عن الآخر، ونظرا لتعقيدات النصوص العربية المكتوبة بخط اليد فلا بد من استخدام العديد من الميزات، وهذه الميزات تنقسم إلى نوعين هما الميزات الكمية (quantitative)، وتشتمل على عدد النقاط وارتفاع الحرف وعرضه ومساحته ووزنه فوق خط الأساس وتحته. أما الميزات النوعية (Qualitative) فإنها تشتمل على الحلقات والتفرعات والمساحات الفراغية ومكان النقطة ونقاط التقاطع والاتصال. في هذا البحث اعتُمد عدد من الميزات، هي: نسب الطول والعرض والارتفاع، والكثافة، والمكونات الثانوية، والحلقات المغلقة، والدوائر المليئة، والتجاويف، واتجاه حركة القلم.

وفي الخطوة الأخيرة قُسّمت الميزات المستخرجة من الأحرف غير المعروفة وجرت مطابقتها مع ميزات مجموعة من الأحرف التي سبق تدريب النظام عليها.

الخطوة الثالثة هي التعرف على ميزات الحرف عن طريق مطابقة ميزاته بالميزات التي تدرب النظام عليها، وهكذا فإن الميزات التي اختيرت في هذا البحث كانت كافية لتحقيق معدل تعرف جيد لجميع الأرقام من صفر إلى ١٠ و(٢٨) حرفا منفصلاً، فبعد استخلاص الميزات لكل حرف دُرب النظام وأُجريت التجارب على عينات واقعية تكونت من مجموعة من الحروف العربية المكتوبة بخط اليد كتبها (٤٨) شخصاً قاموا بكتابة النصوص نفسها، ثم قُيّمت دقة الميزات المختارة، وأثبتت النتائج كفاءة الطريقة المقترحة حيث حققت دقة في التعرف على الحروف قاربت ٧٠٪ بينما حققت نتائج دقة بلغت ٨٨٪ في التعرف على الأرقام.

Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 1, January 2012

Year of publication: 2012

OCR Post-processing Error Correction Algorithm Using Google's Online Spelling Suggestion

Youssef Bassil, Mohammad Alwani

تقنية التعرف الضوئي على الحروف بعد معالجة تصحيح الأخطاء باستخدام الاقتراحات الإملائية لجوجل على الإنترنت

أجهزة الحاسوب الحالية أجهزة إلكترونية ورقمية، وبالتالي بإمكانها معالجة البيانات ذات الشكل الرقمي فقط، لذلك، فإنه لابد من تحويل أي مواد تتطلب معالجة الحاسوب إلى الشكل الرقمي كخطوة أولى. التعرف الضوئي على الحروف(OCR) هو عملية تحويل الصور الممسوحة ضوئياً من النص إلى الوثائق الرقمية القابلة للتحرير التي يمكن معالجتها، وتحريرها، والبحث عنها، وحفظها، ونسخها لعدد غير محدود من المرات دون أي تراجع أو فقدان للمعلومات باستخدام جهاز الحاسوب. على الرغم من أن التعرف الضوئي على الحروف يبدو مثالياً لتحويل المكتبة التقليدية إلى مكتبة إلكترونية، فإنه يعاني من أخطاء وقصور. عملياً، فإن معدل خطأ أنظمة التعرف الضوئي على الحروف يمكن أن يصبح مرتفعاً إلى حد ما، وأحياناً قريباً من ١٠٪. وعندما يفشل نظام التعرف الضوئي على الحروف في التعرف على حرف ما، يتسبب عادة في إحداث أخطاء إملائية في النص الناتج. ولمعالجة هذه المشكلة، يمكن مراجعة نص التعرف الضوئي على الحروف وتصحيحه يدوياً. إلى حد ما، يُعدُّ هذا الإجراء مكلفاً، ومستهلكاً للوقت، وعرضة للخطأ لأنّ العين البشرية قد تفوتها بعض الأخطاء. يمكن استخدام نهج أفضل يقوم على أتمتة تصحيح الكلمات التي تقع فيها أخطاء إملائية باستخدام برامج الحاسوب مثل المدقق الإملائي. يتكوّن هذا الحل من استخدام قاموس البحث للبحث عن الكلمات التي تقع فيها أخطاء إملائية وتصحيحها بشكل مناسب. هذه التقنية لديها أيضاً عيوبها. لهذا السبب

اقتُرحت تقنيات تصحيح الأخطاء القائمة على السياق اللغوي للكشف عن أخطاء التعرف الضوئي على الحروف وتصحيحها فيما يتعلق بسياقها النحوي والدلالي، مما يحسّن كثيرا من معدلات تصحيح أخطاء التعرف الضوئي على الحروف.

للطرق المذكورة أعلاه عيب مشترك هو أنها جميعها تتطلب دمج قاموس واسع من المصطلحات الضخمة التي تغطي تقريباً كل كلمة في اللغة الهدف، كما يجب تحديث محتوي هذا القاموس باستمرار، وبما أنه من المستحيل عملياً تجميع هذا القاموس الواسع، سيكون من الحكمة استخدام محرك بحث جوجل الذي يحتوي على جميع الكلمات والمصطلحات والتعبيرات الممكنة الموجودة في اللغة تقريبا.

تقترح هذه الورقة طريقة جديدة لمعالجة ما بعد تصحيح أخطاء التعرف الضوئي على الحروف استناداً إلى اقتراح إملائي وميزة "هل كنت تقصد" لمحرك بحث في الشابكة بواسطة جوجل. الهدف من هذا النهج هو أتمتة التدقيق اللغوي لنص التعرف الضوئي على الحروف وتوفير الكشف القائم على السياق وتصحيح أخطاء التعرف الضوئي على الحروف. أظهرت التجارب التي أُجريت تحسناً كبيراً في معدل تصحيح أخطاء التعرف الضوئي على الحروف، حيث اكتُشف وصُحح عدد أكبر من الأخطاء الإملائية والأخطاء اللغوية باستخدام الطريقة المقترحة مقارنة بطرق أخرى تقليدية قائمة.

٣-٤-٣ التعرف على أسماء الأشياء

وتضم عشرين بحثا، بينها بحثان من نوع أ، هما: استقصاء أثر تجذيع الكلمات على التعُرّف على الكيانات العربية المسمّاة، ونظام «RENAR» : نظام التعرف على أسماء الأشياء المعتمد على القواعد. وهناك ثمانية عشر بحثا من نوع ب، هي: دراسة مقارنة لتعلّم الآلة للتعرف على الأسماء العربيّة للأشياء، والتعلّم الموجه بالاستدعاء لأنماط الأسماء في الويكيبيديا العربيّة، و طريقة مهجنة لاستخراج العلاقات بين أنماط الأسماء العربيّة، ومنهجية جديدة للتعرف على أنماط الأسماء العربيّة، والتعلّم الموجه بالاستدعاء لأنماط الأسماء في الويكيبيديا العربيّة، وطريقة

مهجنة لاستخراج العلاقات بين أنماط الأسماء العربيّة، ومنهجية جـديدة للتعـرف على أنماط الأسماء العربيّة، و طريقة «ANERsys» للتعرف على أسماء الأشياء باللغة العربيّة مستندة إلى أعلى فوضى، وتنسيق متتال في التعرف على أنماط الأسماء العربيّة من خلال استخدام منهج هجين، وفكّ الغموض عن الأسماء العربيّة باستخدام البيانات المفتوحة المترابطة، والتعرف على أسماء الأشياء العربيّة باستعمال المحول التتابعي وموسوعة ويكيبيديا العربيّة، والتعرف على أسماء الأعلام العربيّة في نصوص مختلفة، والجمع بين الأساليب الخاضعة للحد الأدنى من الإشراف للتعرف على أسماء الأشياء بالعربيّة، و أنظمة التعرف الآلي على أسماء الأشياء في العربيّة باستعمال تقاطع النطاقات، و إضافة قواعد دلالية لتحسين أنظمة التعرف الآلي على أسماء الأشياء، والتعرف على أسماء الأشياء في أسماء الأشخاص في التغريدات العربيّة، وتحسين التغطية والأداء لنظام التعرف على أسماء الأشياء بالاعتماد على القواعد في اللغة العربيّة، ودعم التعرف على أسماء الأشياء باستعمال مؤشرات لا تعتمد على اللغة ومؤشرات خاصة، و استخدام الويكيبيديا موردا للتعرف على أسماء الأعلام العربيّة.

International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 7,No. 1, January 2016

Year of publication: 2016

Exploring The Effects of Stemming on Arabic Named Entity Recognition

Ismail El bazi, Nabil Laachfoubi

استقصاء أثر تجذيع الكلمات على التعرّف على الكيانات العربيّة المُسمّاة

المُقدمة

إيجاد جُذوع الكلمات أو جُذورها: هي عملية إرجاع الكلمات أو المفردات إلى جذوعها أو جذورها. ونظراً لثراء اللغة العربيّة بالقواعد الصّرفية وتعقيدها، تعتبر هذه العملية جزءاً أساسياً لمعظم مهام معالجة اللغة الطبيعية لهذه اللُّغة. تناول هذا البحث دراسة تأثير أساليب التجذيع المختلفة على مهمّة التعرف على الكيانات (الهيئات) المسماة للغة العربيّة، واستقصاء الأُسس المَوضوعيّة والقيود والاختلافات بين أسلوبي التجذيع الخفيف (**Light Stemming**) والتجذيع القائم على استخراج الجذور (**Root-extraction Stemming**). لقد قيُّمت تَجاربنا اعتماداً على قائمة البيانات القياسية للذخيرة اللغوية (**ANERCorp**)، وكذلك اعتماداً على ذخيرة موسوعة (**AQMAR Arabic Wikipedia Named Entity Corpus**).

مشكلة البحث والمبررات

تهدف مهمة التعرف على الكيانات المسماة إلى تحديد الأسماء الصّحيحة والأسماء المهمّة في النص وتصنيفها إلى مجموعة فئات مُحدّدة مُسبقا تكون محل اهتمام أشخاص أو مُنظمات أو مواقع، إلخ. وتعتبر هذه الخُطوة مُهمّة في إطار المعالجة الأوّلية في العديد من تطبيقات مُعالجة اللُّغة الطبيعيّة، بما في ذلك استرجاع المعلومات والترجمة الآلية والتلخيص أو السُؤال والجواب.

وفي حين تركّزت مُعظم الأعمال في مجال التعرف على الكيانات (الهيئات) المُسماة بشكل أساسي على اللُّغة الإنجليزية، بدأت اللغة العربية –على مدى العقد الماضي– تحظى بزخم واهتمام كبيرين في هذا المجال، مع زيادة توافر الذخائر اللغوية العربية المَشروحة. لكن اللغة العربية تحمل – كلُغة ساميّة – خصائص صرفية واشتقاقية مُعقدة وكثيرة. وتُعزى لهذه الخصائص نُدرة البيانات، ولذلك فهي تحتاج إلى ذخائر لغوية ضخمة جداً لتدريب أو تمرين أنظمة التعرف على الكيانات أو الهيئات العربية المُسماة بالمقارنة مع أنظمة (التعرف على الكيانات أو الهيئات المسماة) الإنجليزية. ومن هنا، وللتغلُّب على هذه العقبة، كان تجذير أو تجذيع الكلمة العربية هو أحد الحُلول المُقترحة.

وفيما يلي المباحث الرئيسية لهذه الدراسة:

أولاً: خلفيّة حول خَصائِص اللغة العربيّة والتّحديات ذات العلاقة بالتعرف على الكيانات (الهيئات) المُسمّاة

تُصنّف اللغة العربيّة حسب استخداماتها إلى ثلاثة أشكال:

العربيّة التقليدية (الكلاسيكية): هي النُسخة الرّسمية من اللغة العربيّة التي استخدمت في شبه الجزيرة العربيّة لأكثر من ١٥٠٠ سنة، وكُتبت بها مُعظم النُصوص الدينيّة العربيّة.

العربيّة القياسية الحديثة: هي اللغة الأساسية المكتوبة في وسائل الإعلام والتعليم، فضلاً عن وسائل الاتصال الرئيسيّة في الخَطابة والإذاعات في جميع البُلدان العربيّة. والاختلافات الرّئيسيّة بين العربيّة التقليدية والعربيّة القِياسية هي أساساً في الأسلوب والمفردات، ولكن من حيث التركيب اللغوي، تعتبران متشابهتين تماما. وهذا هو الشكل مدار البَحث في هذه الدراسة.

اللّهجات المحليّة العربيّة: وهو شكل اللغة المستخدم في الحياة اليوميّة في الاتصالات غير الرسمية، ولا تدرس في المدارس و ليس لها معايير مُوحدة. وهي ترتبط تحديداً بالإقليم، ولا تختلف فقط من بلد إلى بلد عربي آخر، إنما أيضاً من منطقة إلى أخرى داخل نفس البلد.

أما التّحديات والمشاكل التي تُواجه أنظمة التعرف على الكيانات أو الهيئات المسماة العربيّة، فهي:

- غياب الحُروف الكبيرة.
- التّلصيق.
- الأصوات (حُروف العِلّة) القَصيرة (الحركات الاختيارية).
- الغُموض الكامن (المتأصِل) في الكيانات أو الهيئات المسماة.
- الأشكال المختلفة للتهجِئة.

ثانياً: استعراض للأنواع المُختَلفة من محللات الجذور العربيّة المُستخدَمة في هذه الدراسة

طُوِّرت العديد من مُحللات الجذور للغة العربيّة، ويمكن تقسيمها إلى نوعين؛ النوع الأول: مُحللات الجذور الخفيفة، وتعتمد على إزالة الإضافات (أي البادئات واللواحق) من الكلمات، في حين يُسمّى النوع الثاني: محللات الجذور القائمة على استخراج الجذور التي تستخرج جُذور الكلمات.

وفي هذا البحث، قدّمنا وصفاً مُختصراً لمختلف محلّلات الجذور أو الجُذوع المستخدمة في هذه الورقة، وهي بالتسلسل:

مُحلّل KHOJA

مُحلّل ISRI

مُحلّل Tashaphyne

محلّل Motaz

مُحلّل Larkey الخفيف

ثالثاً: وصف الإعدادات التّجريبيّة:

يستند نظامُنا (للتعرف على الكيانات المسماة) إلى أسلوب «الحُقول العشوائية المشروطة» (**random fields conditional**). ويُعدُّ العديد من المؤلفين هذا الأسلوب واحداً من أكثر الخوارزميّات مُنافسةً للتعرف على الكيانات المسماة. وقُمنا باستخدام قائمة السّمات التالية في تجاربنا:

الكلمة أو المفردة: الكلمات المحيطة او المجاورة لنافذة السياق (−١، ....، +١).

الجذر: الجذور المحيطة أو المجاورة لنافذة السياق (**context window**) (−١،...، +١).

الإضافات: استخدمت بادئات ولواحق بالجذور، حيث تراوح مدى طولها من (١ إلى ٤).

مقياس (Character n-grams) للحُروف.

الذخائر اللغوية: اعتُمِد على اثنين من الذخائر اللغوية (قوائم البيانات) ، وهما: **ANERCorp** و **AQMAR**. الأول متعلق بوكالات الأنباء، ويشتمل على أكثر من ١٥٠٠٠٠ كلمة مشروحة، والثاني يشتمل على ٧٤٠٠٠ متن نموذجي مأخوذة من ٢٨ مقالة عربيّة على موسوعة ويكيبيديا العربيّة.

الأدوات: استخدمنا في هذا العمل الأدوات التالية:

• **CRF++**، وهي مجموعة أدوات توصيف تسلسل (**sequence labeling toolkit**) مستخدمة مع المعلمات الافتراضية (**default parameters**).

POS tagger، word

.segmenter، normalizer، and a punctuation and diacritic remover

• **AraNLP**، وهي مكتبة قائمة على لُغة برمجة**Java** لمعالجة النُصوص العربيّة. وتتضمّن هذه المكتبة كاشفاً للجُمل (**sentence detector**)، ومُحللاً للقواعد الدلالية، ومحلل جذور خفيفا،

ومُحلل جذور قائما على استخراج الجذور، وأداة وسم أجزاء الكلام **POS tagger** ومُجزّئ الكَلمات لمقاطع (**segmenter**)، ومطبِّع النصوص العربيّة (**normalizer**)، ومُزيل علامات التّرقيم (**diacritic remover**).

• **SAFAR**، وهي منصّة متكاملة تجمع بين جميع طبقات معالجة اللغة الطبيعية العربيّة، وتتضمن هذه المنصّة: مُعالجاً للُّغة الطّبيعية، ومُجزّئًا للجُمل، ومحلّلا لغويا، ومُحلل جذور، ومُحللا نَحويا ومُحللات صَرفيّة.

مقاييس التقييم:

تبنّينا في هذه الدراسة مقياس (**CoNLL**) للتقييم، ويأخذ هذا المقياس الصّارم بالاعتبار الكيانات (الهيئات) الموسومة الصحيحة فقط إذا كانت متوافقة مع الهيئة المتصلة بها في البيانات الذهبية (**gold data**)، وهي مبنية على معادلة الاسترجاع الشهيرة واختبار-**F**.

رابعاً: نتائج التّجارب أو الاختبارات

أظهرت نتائج تجربتنا الأولى أنه حتى أبسط الطُرق يُمكنها تحسين النتائج على كلٍّ من قوائم البيانات (الذخائر لغوية) بالمقارنة مع خط الأساس (**baseline**) القائم على الكلمة. وقد أظهرت النتائج تفوق تقنيات التجذيع الخفيف (**Light Stemming**) بشكل كبير على الأساليب القائمة على تقنيات استخراج الجذر. وتحققت أفضل النتائج على قائمة بيانات الذخيرة اللغوية **ANERCorp** باستخدام محلل الجُذور الخفيف**Light1**. أما بالنّسبة لقائمة بيانات الذخيرة **AQMAR** فقد تفوق المحلل **Light2** بشكل طفيف عند استخدامه مقارنة بالمحلل **Light1**.

أما بالنسبة لنتائج تجربتنا الثانية فقد أظهرت بأن جميع توليفات محللات الجذور أدت إلى تحسين النتائج لجميع قوائم بيانات الذخائر اللغوية بالمقارنة مع المحلل **Light1** (خط الأساس). وتحققت أفضل النتائج على ذخيرة بيانات **ANERCorp** باستخدام مزيج من المحلل **Light1** والمحلل **Tashaphyne**. أما بالنسبة لقائمة بيانات **AQMAR**، فقد تحققت أفضل النتائج من

خلال الجمع بين Light1 و Light8.

وبشكل عام، ووفقاً للنتائج التي حصلنا عليها من جميع تجاربنا، أدّى إدخال الجذر أو الجذع سِمةً إلى تحسين أداء أنظمة التعرف على الهيئات المسماة خاصة باستخدام أسلوب بسيط مثل (aka light stemming). كذلك، يبدو أن الجمع بين الأساليب المختلفة لتحليل الجذور سوف يعزز أداء أنظمة التعرف على الهيئات المسماة العربيّة.

الخاتمة

قمنا في هذه الدراسة باختبار تسعة أساليب مُختلفة لأنظمة (التعرف على الكيانات أو الهيئات المسماة) العربيّة، اعتماداً على قائمتي بيانات ذخيرتين لغويتين هما: (ANERCorp)، و(AQMAR Arabic Wikipedia Named Entity Corpus). ومن ضمنها: مُحلّلا التجذيع الخفيف (Light Stemming) والتجذيع القائم على استخراج الجذور (Root-extraction Stemming). وأظهرت النتائج أن أسلوب التجذيع الخفيف (Light Stemming) يتفوق بشكل كبير على أسلوب التجذيع القائم على استخراج الجُذور (Root-extraction Stemming). وكانت جميع أساليب التجذيع أفضل من خط الأساس القائم على الكلمة. في حين كانت أفضل النتائج التي حُصِّلت باستخدام المحلّل (Light١) على قائمة بيانات (ANERCorp)، حيث بلغت قيمة (مقياس F) له (١ , ٦٩٪). أما بالنسبة لقائمة بيانات الذخيرة (AQMAR)، فقد تحققت أفضل النتائج باستخدام المحلّل (Light٢) بقيمة (مقياس F) بلغت (٠٣ , ٥٧٪)، كما أظهرت النتائج أن الجمع بين الأساليب المختلفة للتجذيع يعزز الأداء العام لأنظمة اللغة العربيّة.

ACM Transactions on Asian Language Information Processing,
Vol. 11, No. 1, Article 2, DOI 10.1145/2090176.2090178
http://doi.acm.org/10.1145/2090176.2090178

Year of publication: 2012

RENAR: A Rule-Based Arabic Named Entity Recognition System

Wajdi Zaghouani

نظام "RENAR" نظام التعرف على أسماء الأشياء العربي المعتمد على القواعد

المقدّمة

لعبت تقنيات التعرف على الكيانات أو الهيئات المسماة دوراً كبيراً في عملية معالجة اللغات الطبيعية، كما هو الحال في استرجاع المعلومات، والتّرجمة الآلية، وأنظمة الإجابة أو الردّ الآلي على الأسئلة. وقد عالج العديد من الباحثين مسألة تحديد الاسم في عدد من اللغات، وقد بدأت بعض الجهود البحثية بالتركيز مؤخّراً على تقنيات التّعرف على الكيانات في اللغة العربيّة. نقدم من خلال هذه الورقة نظاماً مقترحاً لاستخراج المعلومات المدوّنة باللغة العربيّة، يُستخدم لتحليل كمّ هائل من النُصوص الإخبارية في كل يوم لاستخراج أنواع الكيانات المسمّاة للأشخاص والمنظمات والمواقع والتواريخ والأعداد أو الأرقام وكذلك الاقتباسات (الخطاب المباشر). ولم يُطوَّر نظام التعرف على الكيانات للغة العربيّة أصلاً، بل كان يستهدف عدّة لغات غير العربيّة، ثم كُيِّف فيما بعد ذلك لتغطية اللغة العربيّة أيضاً. وتختلف اللغة العربيّة كغيرها من اللغات السامية اختلافاً كبيراً عن اللُّغات الهندو-أوروبية والفنلندية-الأوغرية التي يغطيها النظام حالياً. ومن هنا، يصف هذه البحث الموارد الخاصة باللغة العربيّة التي يتعيّن تطويرها وما هي التّغييرات الواجب إدخالها على القاعدة الموضوعة لكي تكون قابلة للتطبيق على اللغة العربيّة. وبشكل عام، جاءت نتائج التقييم مُرضية، غير أنه من الممكن تحسينها لأنواع مُعينة من الكيانات (الهيئات).

مشكلة البحث

تُستخدم تقنيات التعرف على الكيانات في قطاعات مختلفة في مجال معالجة اللغات الطبيعية واسترجاع المعلومات. لكن معظم هذه الأنظمة أحاديّة اللُّغة (عادة باللغة الإنجليزية). لكن بالمقابل، تغطي متصفّحات الأخبار (NewsExplorer) حاليا ١٩ لُغة، من ضمنها اللغة العربيّة. ومن المرجّح ألاّ تتحقّق مثل هذه الدرجة العالية من التّعدّدية اللُّغوية (multilinguality) إلاّ إذا كانت الجُهود المبذولة لكُل لُغة محددة. ومن المعلوم أنّ اللغة العربيّة تختلف اختلافاً كبيراً عن غيرها من اللغات، وبالتالي، تُعتبر حالة جيّدة لاختبار الطريقة المقترحة في هذا البحث.

اشتملت هذه الورقة على المباحث الرّئيسية التالية:

بعض التّحديات التي تواجه اللغة العربيّة

تناولنا في نظامنا "رينار RENAR" العديد من التّحديات التي تفرضها خُصوصيّات اللغة العربيّة، التي تختلف كثيراً عن اللُّغات الأوروبية المدعومة من مُتصفّح الأخبار (EMM-NewsExplorer). وفيما يلي بعض القضايا الواجب مراعاتها عند بناء نظام للتعرف على الهيئات المسماة باللغة العربيّة.

الغموض وعدم نُطق بعض الحُروف في النّص العربي المكتوب، وهو أمر شائع في مقالات الأخبار.

غياب الحروف الكبيرة من النصوص العربيّة على عكس اللغات اللاتينيّة.

تعقيد قواعد الصّرف العربيّة التي تعتمد على جذور الكلمات وكثرة اشتقاقاتها.

نقص التوحيد القياسي لتهجئة الحروف العربيّة.

هيكل بناء النظام أو الأسلوب المقترح

من أجل بناء نظام رينار استخدمنا الذخائر اللغوية المتاحة بشكل مجاني، لأننا لم نتمكن من ترخيص أيّ من الذخائر اللغوية المتاحة بشكل تجاري. كما أننا قمنا ببناء قائمة من الكلمات المستبعدة، وقائمة المعدلات، والفهارس الجُغرافية لأسماء الأشخاص والمواقع والمنظمات.

الذخيرة اللغوية: قمنا باستخدام الذخيرة (ArabiCorpus) المتاحة بشكل مجاني ذخيرةً رئيسيةً لبناء مواردنا اللاّزمة، وهي نص عربي يشتمل على (٦٨٩٤٣٤٤٧ كلمة) من موارد مختلفة.

الكلمات أو المفردات المحفّزة (Trigger Words): تحاط الأسماء الصّحيحة عادةً بكلمات محفزة أو قرينة.

قائمة كلمات الوَقْف (Word Stop): كلمات الوقف هي كلمات متكررة لا يمكن أن تكون جزءاً من الكيانات المسماة.

الفهارس الجغرافية (Gazetteers): واشتملت على ١٩٦٠٠ اسماً لأشخاص؛ و٢٢٠٠ اسم للبلدان والمدن والبلدات والقرى والولايات؛ وأخيراً ٤٠٠٠ اسم للمُنظمّات والشّركات.

وصف بناء نظام التعرف على الكيانات المسماة العربيّة

بناء النظام: يعتمد النظام على ثلاث خطوات معالجة رئيسية، وهي المعالجة المسبقة (قواعد التقسيم أو التجزئة)؛ والبحث عن الأسماء المعروفة تماماً؛ والتعرف على الأسماء غير المعروفة باستخدام القواعد المحلية؛ وقائمة أو مجموعة القواميس.

المُعالجة الصّرفية المُسبَقة: تمتاز قواعد الصّرف العربيّة بأنها مُعقّدة نسبياً، من حيث إنها تستخدم البادئات واللاّحقات. ويمكن التعامل مع هذا الاختلاف الصّرفي باستخدام القواعد اليدوية لتجزئة السوابق المحتملة واللواحق من جذر الكلمة قبل تطبيق نظام التعرف على الكيانات المسماة لقواعد النحو.

القواعد المحلية للتعرف على أسماء الأشخاص والمنظمات: حيث توضح أولاً كيفية عمل قواعد EMM متعددة اللغات لعمل نظام التعرف على الكيانات المسماة بشكل عام، ثم توصف الاختلافات الخاصة باللغة العربيّة.

نتائج تقييم الأسلوب المقترح

الذخيرة اللغوية: تم تقييم نتائج نظامنا مقارنة بذخيرة التقييم الموحد: (Benajiba

AnerCorp)، والمتاحة مجاناً على الشّابكة. وتتكون الذخيرة من ٣١٦ مقالة و ١٥٠٢٨٦ رمزاً جُمعت من موارد مُختلفة للأخبار على الشّابكة. وتمثل الأسماء الصحيحة ما نسبته ١١٪ من مجموع الذخيرة.

النتائج: قُمنا باستخدام أداة التقييم (CoNLL 2002)، لمعالجة النتائج التي جرى الحصول عليها من خلال نظام رينار. ووفقاً للمبادئ التوجيهيّة الرسمية لأداة (CoNLL)، يُعدُّ الكائن المسمّى قد كُشف عليه بشكل صحيح فقط إذا كان التصنيف صحيحاً، وإذا جرى التعرف على جميع الكلمات المكونة للكيان المسمى. وقد استخدمنا مقاييس التقييم القياسية (أي الدقة، والاستدعاء أو الاسترجاع، ومقياس-F)، مما يُتيح المقارنة المباشرة مع نتائج الأنظمة الأخرى.

مُلخص النتائج

أظهرت النتائج بوضوح تفاوت أداء مختلف الأنظمة تفاوتاً كبيرا حسب نوع الكيان، وكانت درجة دقة الأنظمة المختلفة لكل نوع قريبة نسبيا من بعضها، وخاصة درجات الأنظمة التالية: Renar و Lingpipe و ANERsys 2.0.

في نظامنا، أمكن تحديد أسماء الأشخاص بشكل صحيح في ١٨ ,٧١٪ من الحالات، وهي نتيجة أفضل مقارنة بالأنظمة الثلاثة الأخرى (٤ ,٦٣٪ لـ Lingpipe و ٢١ ,٥٤٪ و ٢٧ ,٥٦٪ لأنظمة Benajiba). أما بالنسبة لفئة الموقع، فقد حصل Renar على (درجة «F" قيمتها ٦٣ ,٨٧٪) وهو ما يفسّر التّغطية الجيدة لنظام رينار وسلامة أسلوب البحث. أما من حيث أسماء الأشخاص والمنظمات، فقد احتلّ نظامنا المركز الثاني (حيث بلغت قيمة F = ٦١,٥٤٪ و٥٢,٢٣٪) على التوالي. ورغم أن هذه القيم تبدو منخفضة إنه عند مقارنتها مع الأنظمة الأخرى يظهر بأنّ جميع الأنظمة التي اختُبرت كانت لديها صعوبات في التعامل مع هاتين الفئتين. أما من جهة الدقة في التعرف على فئة أسماء الأشخاص، فقد أحرز نظامنا المرتبة الأولى، حيث حصل رينار على نسبة (١٨ ,٧١٪) وهي أعلى بست نقاط من Lingpipe. ومن جهة أخرى، كان التحدي الأكبر يتمثّل في فئة أسماء المنظمات، حيث تراوحت قيمة مقياس F

ما بين ٣٦,٧٩٪ لنظام ((ANERsys ٠,٢ و ٢٣,٥٢٪ لنظام (Renar)، و٥,٥٦٪ لنظام Lingpipe)) .

الخاتمة

استعرض هذا البحث جهودنا لتكييف نظام التعرف على الكيانات أو الهيئات المسماة متعدّد اللغات مع اللغة العربيّة. فالعديد من القواعد المستقلّة لُغوياً الموجودة حالياً بحاجة للتكيف مع اللغة العربيّة، ويرجع ذلك في الغالب إلى عدم وجود الحالة الإملائية (orthographic case) الأمر الذي يجعل من الصّعب معرفة أين يبدأ الاسم وأين ينتهي. من جهة أخرى، نحن نستخدم حالياً في مُتصفّحات الأخبار، في الغالب، قواعد تعرف آمنة (مثل تلك التي تستخدم قوائم من أجزاء الأسماء المعروفة والحروف المزيدة للأسماء).

إلا أننا نخطط حالياً للعمل على تحسين عملية الاستدعاء، كما أننا أظهرنا أن هناك بعض العوامل المهمة التي أثرت بشكل كبير على النتائج التي تحققت أعلاه مثل النقص النسبي في الإملاء المُوحد (القياسي) في اللغة العربيّة، خاصةً في نسخ الأسماء الأجنبية، إضافةً إلى لـلغُموض الكبير الذي يكتنف النّص العربي والحَجم الصّغير نسبياً من الفهارس الجغرافية المتاحة لدينا. لكن يمكن القول بأن النتائج التي حصلنا عليها تُعدُّ جيدة نسبياً، بالنظر إلى بساطة الأسلوب وضمان أن يبقى الأسلوب نفسه عبر جميع اللُّغات العشرين التي نتعرّف حالياً على الكيانات المسماة من خلالها.

International Journal on Advanced Science, Engineering and Information Technology,(2)7,p.p. 511-518

Year of publication: 2017

A Comparative Review of Machine Learning for Arabic Named Entity Recognition

Ramzi Esmail Salah, Lailatul Qadri binti Zakaria

دراسة مقارنة لتعلم الآلة للتعرف على الأسماء العربيّة للأشياء

الملخص:

تهدف أنظمة التعرف على الأسماء العربيّة للأشياء (Arabic named entity recognition) إلى التعرف على أسماء الأشياء (name entity) وتصنيفها بالعربيّة في النص العربي، وهناك مهام مهمة في معالجة اللغة الطبيعية (natural processing language) العربيّة تحتاج التعرف على أسماء الأشياء في الترجمة الآلية (machine translating)، سؤال وجواب (question-answering)، استخراج المعلومات (information extraction)، الخ.

وبشكل عام فإنّ أنظمة التعرف على أسماء الأشياء في العربيّة يمكن أن تصنف إلى طريقتين: حسب القواعد (rule-based)، تعليم الآلة أو نظم هجينة (machine learning or hybrid systems). وفي هذه الورقة دُرِس تعليم الآلة والتعرف على أسماء الأشياء العربيّة ومقارنتها بالمصادر اللغوية ونوع الكائن (entity)، والمجال (domain)، والطريقة (method) والأداء (performance).

المقدمة:

تُعدُّ تقنية التعرف على أسماء الأشياء عند استخدامها كخطوة تمهيدية أو استباقية لتطبيقات معالجة اللغة الطبيعية مؤشراً مهماً في تحسين كفاءة الأداء الكلي لها. ومع ذلك، من الممكن أن

يتكون النص من نوع واحد أو أكثر من الأسماء، كأسماء الأشخاص أو المواقع، أو المؤسسات أو المواقع الرياضية، والعديد من الأسماء الأخرى من مجالات (domains) وحقول معينة. هذه الأسماء تسمى أسماء الأشياء، والتعرف الآلي على أسماء الأشياء وجد لتعريف هذه الأسماء وتصنيفها آلياً من النصوص إلى طبقات معرفة مسبقاً (predefined classes)، وهو مشروع مهم في العشر سنوات الأخيرة في هذا المجال لا سيما في مجال اللغة العربيّة. والنظام المطروح هنا استخدم طرقاً متنوعة مختلفة وتقنيات عديدة للتعرف على أسماء الأشياء، و كان التعلم الآلي (machine learning) الطريقة الفضلى والتقنية الأكثر نجاحا نظراً لقدرتها على استيعاب مجالات لغوية متعددة وسهولة تجربتها واستعمالها. وفي هذه الدراسة أيضاً استخدمنا هذه النظم في التقنية المستخدمة للتعرف على أسماء الأشياء في العربيّة (ANER) وقدمنا ملخصاً للتقارير التي تضمنت نوع اللغة والنطاق، نوع الكينونة (entity type)، الطريقة والأداء (الإنجاز performance)، كما قدمت هذه الدراسة أيضاً النماذج (models) ومزايا التعرف على أسماء الأشياء (NER features) المستخدمة في منهج التعلم الآلي (Machine Learning) مع بعض التفاصيل للتحديات المتعلقة بتقنية التعرف الآلي على أسماء الأشياء في العربيّة.

النتائج والمناقشة:

خلال العقد الأخير طور الباحثون أنظمة ترجمة آلية للتعرف على أسماء الأشياء في العربية باستخدام نماذج مخصصة للتعليم الآلي (machine learning)، وقد أجريت العديد من البحوث على دراسات في مجال الترجمة الآلية والتعرف على أسماء الأشياء الخاضع للإشراف مع اهتمام قليل بالنوع شبه الإشرافي، ولم يُعد تقرير خاص بالنوع غير القابل للإشراف (unsupervised). كما أن الدراسات في مجال التعرف على أسماء الأشياء للنصوص العربيّة الحديثة (modern standard Arabic) ركزت على قليل من أنواع أسماء الأشياء وقليل من الحقول بينما نادراً ما تمّ التحقق من الحقول الأخرى.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, pages 162-173, Avignon, France.

Year of publication: 2012

Recall-Oriented Learning of Named Entities in ArabicWikipedia

Behrang Mohit, Nathan Schneidery Rishav Bhowmick, Kemal Oflazer, Noah A. Smithy

التعلم الموجه بالاستدعاء لأنماط الأسماء في الويكيبيديا العربيّة

تُعدُّ مشكلة التعرف على أنماط الأسماء من المشاكل الأساسية في موسوعة الويكيبيديا العربيّة، حيث لا يوجد بيانات تدريب كافية في هذا المجال، فتركزت معظم الأبحاث السابقة على دراسة التعرف على أنماط الأسماء في إطار نصوص المقالات الإخبارية فقط، مما أدى إلى ضعف تطبيق استدعاء الأسماء في المجالات الأخرى مثل الويكيبيديا، وصعوبة التعرف على أنماط الأسماء فيها. لذلك يحاول الباحثون تدريب أنظمتهم على التكيف مع التعلم الذاتي على بيانات النطاق المستهدف غير المدربة (untrained).

تتناول هذه الورقة البحثية مشكلة التعرف على أنماط الأسماء في نصوص مختلفة، وهذا ما يميزها عن معظم البحوث السابقة في هذا المجال. وعلى وجه التحديد، أخذ الباحثون بعين الاعتبار بعضا من مقالات الويكيبيديا العربيّة مع مواضيع متنوعة خارج نطاق الأخبار شائعة الاستخدام. ولم تكن هذه بالمهمة سهلةً على الباحثين، حيث إن اللغة العربيّة هي لغة غنية من الناحية الصرفية، وتؤدي هذه السمة مواجهة إلى عدة عقبات في هذا المجال، أهمها: أن الكتابة العربيّة الشائعة خالية من التشكيل (الفتحة والضمة والكسرة)، وهذا يزيد من الغموض المعجمي، إضافة إلى أنها لغة تفتقر إلى وجود الأحرف الكبيرة كما هو الحال في بعض اللغات الأخرى مثل الإنجليزية. أما العقبة الثانية فتكمن في قلة البيانات المتوفرة في هذا المجال، فكان

الباحثون يركزون اهتمامهم على المقالات الإخبارية فقط دون المجالات الأخرى، وبالطبع أدى ذلك إلى شح الموارد المتوفرة لدى الباحثين للاعتماد عليها في دراساتهم. فقد هدف الباحثون إلى توفير نماذج اختبار جديدة لتقييم الأنواع المختلفة من أنماط الأسماء (أسماء الأشخاص، الأماكن، المنظمات وغيرها)، وذلك عن طريق إنشاء مجموعة بيانات كأمثلة للتطوير والاختبار، ولكن ليس كبيانات تدريبية.

وقد ساهم الباحثون في هذا العمل بعدة مجالات: أولا بناء مجموعة صغيرة من بيانات التدريب خارج نطاق البيانات الإخبارية. ثانيا تعزيز دور منهج خوارزميات تعلم الآلة وكذلك التعلم الذاتي. ثالثا إجراء تقييم لأداء هذا النظام بناء على نموذج يحقق لهم الغرض المستهدف، حيث أظهرت التجارب الأولية نتائج جيدة لحل مشكلة تحديد أنماط الأسماء في ويكيبيديا. لقد اقترح الباحثون حلا لمشكلة التعرف على أنماط الأسماء العربيّة لمختلف النطاقات التي تفتقر لوجود بيانات تدريب كافية، فُحصول على بعض البيانات في هذا المجال مع تحسين ملحوظ في مستوى الدقة والأداء للكشف عن النمط، وخصوصا عندما يقترن مع نظام التعلم الآلي (الذي دُرِّب مسبقا على أنماط الأسماء) . كما طوّر الباحثون مجموعة صغيرة من مقالات الويكيبيديا العربيّة من خلال مخطط توضيحي يغطي عدة مجالات ومواضيع.

Journal of King Saud University – Computer and Information Sciences (2014) 26, 425–440

Year of publication: 2014

A hybrid method for extracting relations between Arabic named entities

Ines Boujelben, Salma Jamoussi, Abdelmajid Ben Hamadou

طريقة مهجنة لاستخراج العلاقات بين أنماط الأسماء العربيّة

نظراً للكمية الهائلة من النصوص الإلكترونية العربيّة، يلاحظ أن هناك تكراراً واسعاً في أنماط الأسماء العربيّة التي لا تمتلك روابط علائقية تجمع بينها. يمثل التعرف على هذه الأسماء المهمة الأولى نحو بناء التحليل الدلالي لاستخراج المعلومات واتخاذ القرارات. والمهمة الثانية تتمثل في استخراج العلاقات الدلالية بين أنماط الأسماء التي تكون مفيدة لفهم أفضل اللغات الطبيعية، وبالتالي، فإن المهمة الثانية تشكل خطوة حاسمة نحو تطبيقات معالجة تلك اللغات. ويتيح هذا النوع من المعلومات مهمة اكتشاف علاقة أو تفاعل مفيد بين أنماط الأسماء المختلفة في محتوى النص. وقد تلقى هذا النهج قدراً كبيراً من الاهتمام لأنه يستخدم في العديد من التطبيقات الخاصة بمعالجة اللغة الطبيعية، مثل التلخيص التلقائي والإجابة الآلية عن الأسئلة. في الواقع، يمكن الأستفادة من استخلاص العلاقات بين أنماط الأسماء لاستخراج إجابات أكثر دقة وصحة. ولذلك، فقد أجريت بالفعل عدة دراسات عن التعرف على أنماط الأسماء في العديد من اللغات مثل الإنجليزية والفرنسية والصينية. بالإضافة إلى ذلك، فقد بُنيت العديد من أنظمة التعرف على كيانات أنماط الأسماء العربيّة.

في هذه الورقة البحثية، جمع الباحثون مزايا تقنيات التعلم الآلي وأساليب تستند إلى قواعد لاستخراج العلاقات بين أنماط الأسماء العربيّة بالنظر إلى أن اللغة العربيّة هي لغة صرفية غنية، فقد قام الباحثون ببناء نموذج لغوي وتعليمي للتنبؤ بمواضع الكلمات التي تعبر عن علاقة

دلالية ضمن الجملة. واعتمد الباحثون على أنماط يدوية عندما تكون أمثلة العلاقات المعطاة معقدة أو معبرا عنها من خلال أكثر من كلمة واحدة. وكانت الفكرة الرئيسة هي استخدام وحدات لغوية لتحسين النتائج التي يجري الحصول عليها باستخدام طريقة تعلم الآلة.

لقد حققت هذه الطريقة أداءً مشجعاً، وكانت النتائج التي جرى الحصول عليها مرضية، كما كانت بمثابة تحفيز لاستراتجية الجمع بين كلا النوعين من الطرق لتعزيز الأداء العام للعملية. وأظهر الباحثون تأثير كلّ وحدة لغوية مستخدمة لإنتاج مكاسب كبيرة مقابل نتائج سابقة. وبالمثل، فقد درس الباحثون أيضاً تأثير الجمل الاسمية والفعلية على أداء النظام المقترح. وأخيراً، فإن المزيد من القيود التي أُضيفت إلى القواعد التلقائية التي تولدت من التقنية الجينية المقترحة تعطي نتائج أكثر دقة وإيجازاً. أظهرت النتائج التجريبية أن النهج الهجين تفوق على كلّ من النظام القائم على القواعد بنسبة(١٢ ٪) والمناهج القائمة على التعلم الآلي بنسبة (٩٪) لتحقيق (٢٪ , ٧٥) عند تطبيقها على نفس مجموعة بيانات الاختبار القياسية. خلص الباحثون إلى أن الأسلوب القائم على هذا النمط يقدم قيماً دقيقة جيدة، ولكن يمكن أن تكون ضعيفة عندما تواجه مفردات غير متجانسة وجملاً معقدة. من ناحية أخرى، فإن تقنية التعلم الآلي أكثر كفاءة من حيث تغطيتها لمجموعة البيانات الخاصة بالباحثين.

The International Arab Journal of Information Technology, Vol. 14, No. 3, May 2017

Year of publication: 2017

A New Approach for Arabic Named Entity Recognition

Wahiba Karaa, Thabet Slimani

منهجية جـــديدة للتعـــرف على أنماط الأسماء العربيّة

إن التعرف على أنماط الأسماء مهمة فرعية لا غنى عنها في مختلف النظم التي تتعامل مع المعالجة التلقائية لأية لغة الطبيعية، مثل استخراج المعلومات، والتعرف التلقائي على الكلام، وفهرسة الوثائق، ووضع حواش للوثائق، وتصنيفها، والترجمة الآلية والإجابة الآلية عن الأسئلة، الخ. يقوم التعرف على أنماط الأسماء بدور جدير بالاهتمام في الأبحاث الخاصة بمعالجة اللغات الطبيعية؛ لأنه يوفر الكشف عن الأسماء المناسبة في النصوص غير المنظمة، كما يجعل من السهل البحث عن المعلومات واسترجاعها واستخراجها، فالمعلومات المهمة في النصوص عادة تقع حول الأسماء الصحيحة. لقد ظهر مفهوم التعرف على أنماط الأسماء في عام ١٩٩٠ في وكالة مشاريع البحوث المتقدمة لتطوير أساليب جديدة لاستخراج المعلومات. وتؤثر جودة نظام التعرف على أنماط الأسماء بشكل رئيس على جودة نظام معالجة اللغة الطبيعية بأكمله. كما توجد في الوقت الحالي مجموعة كبيرة ومتنوعة من أدوات التعرف على أنماط الأسماء، وذلك لبعض اللغات المستخدمة بشكل كبير فقط. ولا يزال التعرف على أنماط الأسماء العربيّة يمثل حيزاً بحثياً محدوداً، فما زالت في طورها التمهيدي مقارنة مع لغات أخرى مثل اللغة الإنجليزية.

إن اللغة العربيّة هي لغة سامية تمتاز بالتنوع الصرفي والهجائي المعقد الذي قد يتسبب بتعقيد التعرف على أنماط الأسماء العربيّة. تواجه هذه اللغة بعض التعقيدات التي تشتمل على ما يلي: أولاً: عدم ابتداء الأسماء بأحرف كبيرة، وهو ما يشكل عقبة رئيسة؛ لأن الأحرف الكبيرة

ذات أهمية قصوى في اللغات اللاتينية لتمييز الأسماء. ثانياً: وجود أحرف العلة، فالأسماء التي تحتوي على أحرف العلة تتصف بالغموض في المعنى أو الوظيفة النحوية. ثالثاً: تركيب الكلمات، فاللغة العربيّة هي لغة شديدة التأثر بالإضافات مثل حروف العطف والجر، وضمائر الملكية، والمحددات التي عادة ما تتصل بالكلمات في بدايتها أو في آخرها. ومعظم الكلمات باللغة العربيّة مشتقة من جذور باستخدام أنماط أو قوالب.

في هذا البحث، اقترح الباحثون نظاماً يستخدم مزايا تقنيات التعلم، جنباً إلى جنب مع النماذج الإحصائية لاستخراج العلاقات من شبكة الإنترنت، وتحديد العلاقات الأكثر أهمية للتعرف على أنماط الأسماء باستخدام مقاييس مختلفة لترتيب تلك العلاقات. يمكن للنظام المقترح تحديد أنماط الأسماء العربيّة دون الحاجة إلى التحليل الصرفي، أو النحوي، أو المعاجم اللغوية. إن الهدف من الطريقة المقترحة هو توفير نظام عام للتعرف على أنماط الأسماء العربيّة، حيث يتعلم هذا النظام التعرف التلقائي على أنماط الأسماء العربيّة، ويربط بينها بشكل منهجي. وقد أثبتت نتائج التقييم التجريبي الشامل عن طريق العديد من مقاييس الأداء الشهيرة كفاءة النظام بمعدل إجمالي يساوي (٨٣٪).

Machine Translation. CWMT 2014. Communications in Computer and Information Science, vol 493. Springer, Berlin, Heidelberg

Year of publication: 2014

A Novel Hybrid Approach to Arabic Named Entity Recognition

,Mohamed A. Meselhi, Hitham M. Abo Bakr, Ibrahim Ziedan and Khaled Shaalan

طريقة هجينة جديدة للتعرف الآلي على أسماء الأشياء باللغة العربيّة

تُعدُّ عملية التعرف الآلي على أسماء الأشياء خطوة تمهيدية أساسية للعديد من تطبيقات معالجة اللغة الطبيعية باعتبارها ملخص النصوص (text summarization) ومصنف المستندات (document categorization) واسترجاع البيانات (information retrieval). وتتبع تقنية التعرف الآلي على أسماء الأشياء إما طريقة تعتمد على القواعد (rule-based) أو طريقة التعلم الآلي (machine learning). في هذه الدراسة قدمنا مشروعاً جديداً للتعرف الآلي على أسماء الأشياء باستخدام طريقة هجينة بحيث تدمج بين التعلم الآلي والطريقة المبنية على القواعد، في إطار تحسين أداء نظم التعرف على أسماء الأشياء، ويدعم النظام التعرف على أسماء ثلاثة أشياء بحيث تضم أسماء الأشخاص، والمواقع والمؤسسات، والتجارب العملية، وقد أظهرت أن نظامنا المهجن أنجز بشكل أفضل من استخدام النظام المبني على القواعد ونظام التعلم الآلي عند معالجتهما بشكل منفصل، وتفوقت على (النظام المهجن للتعرف الآلي على أسماء الأشياء بالعربيّة).

تهدف عملية التعرف على أسماء الأشياء إلى تحديد سلسلة من الكلمات في نص معين يمكن تصنيفها تحت تصنيفات مُعرَّفةٍ مسبقاً لأسماء الأشياء. ولأن العربيّة لغة معربة (inflected) لها تركيبتها الصرفية المعقدة الغنية، فقد ركزنا في هذه الورقة بشكل خاص على التعرف الآلي على أسماء الأشياء بالعربيّة، كما قدمنا تكاملاً بين مكون يعتمد على القواعد، وإعادة إنتاج لنموذج

تعرُّف آلي عربي على أسماء الأشياء مع مكون تعلم آلي للتعرف على أسماء الأشياء، وبشكل خاص SVM بهدف تحقيق مزايا وفوائد كلا النظامين وتقليل المشاكل والأخطاء الواردة فيهما.

يعتمد النظام المبني على القواعد على مجموعة من القواعد اللغوية، والنظام المبني على التعلم الآلي يعتمد على مجموعة من المزايا المستخرجة من نص مشروح، مذيل (annotated text).

الخلاصة:

طريقتنا الهجينة في التعرف على أسماء الأشياء كاملت بين الطريقة المبنية على القواعد (rule-based) والتعلم الآلي (machine learning approach) بهدف تحسين الأداء الكلي للنظام. كل من المكونين يستجيبان لنهج متكامل عولجا بالتوازي، أي باختيار العلامة (tag selection) والتصحيح. وقد استخدمت هذه الطريقة لتحسين مخرجات نظام التعلم الآلي من خلال اختيار السلبيات الأكثر خطأً مثل التعليقات – الشروحات المفقودة (missing annotations) وتطبيق التصحيحات باستخدام توسيم القرارات (tagging decisions) التي تحددت من خلال المكون المبني على القواعد (rule-based component).

أظهرت النتائج نسبة ٦٥ ,٩٥٪ و٩ ,٩٢٪ و٨ ,٩٤٪ لكل من أسماء الأشخاص والمباني والأماكن على التوالي، لذا تفوق النظام الهجين في إظهار النتائج، وأشارت دراستنا إلى أن تأثير المزايا بيّن أنه عندما يكون حجم النافذة ٣ يكون الأداء الأفضل. وللعمل المستقبلي نتطلع إلى زيادة قابلية النظام لتعريف الأنواع الأخرى من أسماء الأشياء، كما يقترح شمول تقنيات مختلفة في التعلم الآلي غير SVM ودراسة تأثيرها على الأداء الكلي للنظام المهجن للتعرف على أسماء الأشياء في العربيّة.

Proceedings of COLING 2012: Technical Papers, pages 2159–2176, COLING 2012, Mumbai, December 2012.

Year of publication: 2012

A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach

Mai Mohamed Oudah, Khaled Shaalan

تنسيق متتال في التعرف على أنماط الأسماء العربيّة من خلال استخدام منهج هجين

ازداد الاهتمام في التعرف على أنماط الأسماء في اللغة العربيّة في العقد الماضي فطُوِّرت هذه الأنماط باستخدام نهجين مختلفين: النهج القائم على القواعد النحوية المصنوعة يدويا من اللغويين، وتحتاج إلى مجهود كبير وتستغرق وقتاً طويلاً، خاصة إذا كان اللغويون الذين لديهم المعرفة والخلفية المطلوبة غير متوفرين. والنهج الثاني هو القائم على تعلم الآلة التي تكون قابلة للتحديث مع الحد الأدنى من الوقت والجهد طالما تتوفر مجموعات كبيرة لذلك. التعرف على أنماط الأسماء هو مهمة كشف وتصنيف الأسماء الصحيحة ضمن النصوص إلى أنواع محددة مسبقا، مثل أسماء الأشخاص والمواقع والمنظمات وغيرها، بالإضافة إلى الكشف عن التعبيرات العددية، مثل التاريخ والوقت والسعر ورقم الهاتف.

بُنيت معظم أنظمة التعرف على أنماط الأسماء العربيّة من خلال إقرار منهجية القواعد أو استخدام المنهجية المبنية على تعلم الآلة، بما فيهما من نقاط قوة وضعف. في هذه الورقة البحثية، يُعالج التعرف على أنماط الأسماء في اللغة العربيّة من خلال دمج المنهجيتين معا في تنسيق متتال لتشكيل المنهج الهجين في محاولة لتحسين أداء مهام التعرف على أنماط الأسماء. النظام المقترح قادر على التعرف على أنواع مختلفة من أنماط الأسماء وقد أُجريت تجارب مكثفة باستخدام ثلاثة مصنفات مختلفة تطبق تعلم الآلة لتقييم أداء النظام الهجين.

وتجدر الإشارة إلى أن خط الأساس (base line) في جميع التجارب التي أُجريت هو أداء العنصر القائم على القاعدة. وأظهرت النتائج التجريبية أن تكييف المنهج الهجيني يؤدي إلى أعلى مستوى أداء. ومن الجدير بالذكر أن نتائج النظام المختلط المقترح قريبة جدا من نتائج المكون القائم على القواعد عندما يتعلق الأمر بالتعبيرات العددية والزمنية. وقد حقق النظام المختلط المقترح تحسنا إجماليا لأداء اللغة العربيّة، حيث ظهر أنها قادرة على التعرف على ١١ نوعا مختلفا من أنماط الأسماء بما في ذلك أسماء الأشخاص، والأماكن، والمنظمات، والتواريخ، والأوقات، والأسعار (الأموال)، والمقاييس (المقادير القياسية)، والنسب المئوية، وأرقام الهواتف، والرقم الدولي المعياري للكتاب (IBSN)، وأسماء الملفات.

تظهر النتائج التجريبية تفوق المنهج الهجين على كل من المنهج المبني على القواعد والمنهج المبني على تعلم الآلة كلا على حدة، حيث إنه يتفوق على أفضل الأنظمة المنشورة في الأبحاث العلمية في مجال التعرف على أنماط الأسماء العربيّة من حيث الدقة عند التطبيق بنتيجة معدلات توافقية (F-measure) قدرها : ٤ , ٩٤٪ في حالة أسماء الأشخاص، و١ , ٩٠٪ في حالة أسماء الأماكن و ٢ , ٨٨٪ في حالة أسماء المنظمات.

A. Gelbukh (Ed.): CICLing 2007, LNCS 4394, pp. 143–1532007 ,

Year of publication: 2007

ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy

Yassine Benajiba, Paolo Rosso, and Jos´e Miguel Bened´ı Ruiz

طريقة "ANERsys" للتعرف على أسماء الأشياء باللغة العربيّة مستندة إلى (أعلى فوضى)

سمحت مهمة التعرف الآلي على أسماء الأشياء (Named Entity recognition) بالتعرف على أسماء محددة إضافة إلى التعابير الزمانية والمكانية في نص مفتوح المجال (open domain text)، كما أثبتت أنظمة التعرف الآلي على أسماء الأشياء أنها ذات أهمية بالغة في العديد من مهام معالجة اللغة الطبيعية (Natural processing language) كاستخراج البيانات (information retrieval) ومهام السؤال والجواب (question answering)، ولسوء الحظ فإن الدافع الأساسي لبناء أنظمة تعرف آلي على أسماء الأشياء للغة العربيّة جاءت من منظور تجاري وكانت دقة الطرق المستخدمة غير معلومة.

في هذه الورقة بُني نظام تعرف آلي على أسماء الأشياء حصرياً لنصوص عربيّة تعتمد على (n-grams) وأقصى حالات الفوضى (maximum entropy) أو التشويش، بالإضافة إلى أننا أنتجنا نظاماً عربياً متخصصاً مستقلاً مع نص صحافة (gazatteers) استخدم لتسريع النظام كما قمنا ببناء وتطوير مجموعتنا الخاصة من المجموعة التجريبية ومجموعة الاختبار لتقييم وتسريع التقنية المطبقة. وقد أظهرت جميع التجارب والنتائج أن الطريقة المعروضة عالجت بنجاح مشكلة التعرف الآلي على أسماء الأشياء باللغة العربيّة.

الخلاصة والعمل المستقبلي:

استعرضت هذه الدراسة مشروع **ANERsys**، وهو نظام التعرف الآلي على أسماء الأشياء باللغة العربيّة، وقد وُجِّه بشكل جماعي مع **ANERcorp** و **ANERgazat** وهي الموارد التي طورت في سياق تطبيق النظام.

وفي إطار تنفيذ مهمة التعرف الآلي على أسماء الأشياء وُظفت طريقة الحد الأعلى للعشوائية (**maximum entropy**) حيث أثبتت هذه الطريقة أنها حل مناسب لمهمة التعرف الآلي على أسماء الأشياء، نظراً لاعتمادها على القواعد (**feature based method**)، وساهمت في رفع كفاءة النظام إلى ١٢ نقطة دون استخدام معلومات مُوَسِّم أقسام الكلام (**POS tagger**) أو تقسيم النص (**text segmentation**) كما أجرينا محاولة أخرى لدمج النظام مع نص صحفي يستند إلى الشابكة (**web-based gazatteers**). لكن وجد أن ذلك لم يحسّن من أداء النظام، وذلك لأن حجم المادة الصحفية قليل نسبياً، وقد خططنا لتحقيق المزيد في هذه المسألة.

والفرق الرئيسي الملاحظ بين أسماء المواقع والتسميات الأخرى هو أن جودة النظام وكفاءته ترتبط وتعتمد على ورود هذه الأسماء في المجموعة التدريبية لأن أسماء المواقع يمكن أن تظهر في أكثر من سياق بعكس الأسماء الأخرى، ولهذا السبب خططنا لزيادة المجموعة التدريبية والمجموعة الفاحصة للحصول على أفضل النتائج. كما استخدمنا طريقة (**ad-hoc**) للتعامل مع البيانات الظاهرة وذلك لطبيعة اللغة العربيّة، وخططنا لاستخدام خوارزميات صلدة (**robust**) قبل تجربة النظام لتقطيع النصوص (**text segmentation**)، كما قررنا توسيم أقسام الكلام في مجموعتي الفحص والتدريب لأنها مزية مهمة لإنجاز نظام عالي الجودة ومضمون الكفاءة للتعرف الآلي على أسماء الأشياء.

7th International Conference on Information and Communication Systems (ICICS)

Year of publication: 2016

Arabic Named Entity Disambiguation Using Linked Open Data

Omar Al-Qawasmeh, Mohammad AL-Smadi, Nisreen Fraihat

فكّ الغموض عن الأسماء العربيّة باستخدام البيانات المفتوحة المترابطة

تمثل الشبكة الدلالية (Semantic Web) رؤية جديدة لشبكة الإنترنت التقليدية، حيث توجد الروابط بين الوثائق على شبكة الإنترنت اعتمادا على خصائص ذات مغزى تصف علاقات التشاركية بين الوثائق (linked documents relationships). لقد بدأ مشروع البيانات المفتوحة المرتبطة ((Linked Open Data (LOD) في عام ٢٠٠٧ سعياً لتحقيق الأهداف الرئيسة لرؤية الشبكة الدلالية من خلال ربط البيانات المفتوحة والبيانات المنظمة من مجالات مختلفة. تنمو البيانات المفتوحة المرتبطة بشكل متسارع حيث تُنشر مليارات الأنماط على شبكة الإنترنت ويجري ربطها ببعض. ومع ذلك، يتطلب هذا النمو الهائل استخراج العلاقات بين تلك الأنماط على شبكة الإنترنت وتمثيلها بطريقة قابلة للقراءة آلياً باستخدام أنظمة وصف الموارد، حيث لا يتوفر سوى خمسة مجاميع من مجموعات البيانات المفتوحة المرتبطة باللغة العربيّة.

يُعدُّ الشرح الدلالي للأنماط (Semantic annotation of entities) التي تظهر في وثائق شبكة الإنترنت أحد أبرز التحديات التي تواجه رؤية الشبكة الدلالية. يهدف الشرح الدلالي إلى ربط الأنماط على شبكة الإنترنت مع نظيراتها الممثلة في قواعد المعرفة. وتوجد أنواع مختلفة للأنماط المذكورة على شبكة الإنترنت على النحو التالي: شخص، منظمة، موقع، الجغرافية السياسية، منشأة، مركبة، سلاح، منتج، الخ. ومع ذلك، قد تظهر الأنماط المذكورة على شبكة الإنترنت بمظاهر مختلفة تعتمد على سياق لغة الوثيقة. وفي بعض الحالات يبدو أن النمط ينتمي

إلى أكثر من نوع واحد، الأمر الذي يؤدي إلى حدوث بعض الغموض. لذلك أصبح توضيح أنماط الأسماء ضرورياً. وتعتمد هذه العملية على ربط الاسم المذكور مع النص بهيئته المناسبة في قاعدة المعرفة. ويستند هذا التوضيح إلى سياق الاسم ونوعه. علاوة على ذلك، ينبغي أن يكون للأنماط الصالحة في قاعدة المعرفة رابط URL صالح (معرف الموارد العالمي). ومن المفاهيم التي يجدر الاهتمام بها في هذا الصدد هو التعرف على أنماط الأسماء، حيث يعبر هذا المفهوم عن نسبة الاسم المذكور إلى نوعه، في حين أن توضيح نمط الاسم هو عملية نسبة النمط المذكور إلى تمثيله حسب قاعدة المعرفة.

يهدف هذا البحث إلى معالجة مشكلة توضيح أنماط الأسماء العربيّة المسماة من خلال نهج محسن لاستخلاص معلومات من موسوعة ويكيبيديا العربيّة والبيانات المفتوحة المرتبطة بها. ويستخدم هذا النهج توسيع تصنيف الأستعلام وتقنيات تشابه النص لتوضيح الأنواع التالية من أنماط الأسماء: مثلا اسم شخص أو موقع أو منظمة. وقد أعدت مجموعة البيانات المرجعية لتوضيح أنماط الأسماء العربيّة وشرحها باستخدام أكثر من عشرة آلاف نمط. النهج المقترح هو نهج هجين لتوضيح أنماط الأسماء، حيث يستفيد من قواعد المعرفة مثل الويكيبيديا لتوضيح أنماط الأسماء العربيّة، كما أنه يقدم مجموعة بيانات مرجعية عربيّة لتوضيحها تتألف من (١٠٠٦٨) نمطاً. وفي النهاية أظهرت نتائج الاختبارات أن دقة النهج المقترح تبلغ (٨٤٪) في مجموعة البيانات الكلية، وأن دقته بالنسبة لأسماء الأنماط المتعلقة بالمكان والشخص والمنظمة تبلغ (٩٤٪) و (٧٦٪) و (٧٨٪) على التوالي.

Proceedings of Recent Advances in Natural Language Processing, pages 48–54, Hissar, Bulgaria, Sep 7–9 2015.

Year of publication: 2015

Arabic Named Entity Recognition Process using Transducer Cascade and Arabic Wikipedia

Fatma Ben Mesmia, Nathalie Friburger

التعرف على أسماء الأشياء العربيّة باستعمال المحول التتابعي وموسوعة ويكيبيديا العربيّة

إن المحول التتابعي (Transducer Cascade) هو نموذج شائع الاستخدام في العديد من تطبيقات معالجة اللغات الطبيعية (NLP)، ويلعب دورا مهما في قضايا استخراج المعلومات(IE) والتعرف على الأشياء((Recognition NER(Named Entity)، لا سيما التعرف على أسماء الأعلام العربيّة، ويرمز له اختصارا بـ ANE)). كما أن المحول التتابعي يعتمد على قراءة المعلومات على شكل ملفات نصية ثم معالجتها، وأغلب الأعمال في هذا المجال تعتمد على المقالات التي توفرها موسوعة ويكيبيديا. وموسوعة الويكيبيديا هي مورد معلومات مجاني يتوفر فيه عدد هائل من المقالات مختلفة الموضوعات ومتنوعة المحتوى. والإصدار العربي من هذه الموسوعة في كثير من الأحيان يميز أسماء الأشخاص بشكل خاص وأسماء الأعلام والأشياء بشكل عام، وعلاوة على ذلك فإن هذه الموسوعة تمتاز بتحديث معلوماتها بشكل مستمر.

تُعدُّ مهمة التعرف على أسماء الأشخاص من أكثر المهام صعوبة بالنسبة لتطبيقات معالجة اللغة العربيّة، وذلك لأن هذه المهمة تتطلب استخدام مختلف القواميس، وقائمة من جذور الكلمات لبناء واستخراج القواعد التي تهدف إلى تطوير مجموعة من نماذج التحويل التي تطبق على قواعد بيانات محددة. في هذا السياق، يهدف البحث الحالي إلى استخدام النهج القائم على

القواعد لبناء محول تتابعي للتعرف على أسماء الشخصيات العربيّة.

تمتاز أسماء الشخصيات العربيّة بالتنوع الكبير، وذلك لأنها في الغالب تتعلق بالدولة الأم للشخص، أو الدين، أو الثقافة، أو حتى الصفات الشخصية. ففي هذا البحث بنيت قاعدة بيانات تحوي عددا من الملفات النصية التي جرى الحصول عليها من(١٩) دولة عربيّة وتحتوي (٧٩٦٥٩) كلمة. وقد استخدمت قاعدة البيانات هذه لحوسبة مجموعة من القواعد لاستخدامها لاحقاً في المحول التتابعي. وبشكل عام فإنه يمكن تصنيف أسماء الشخصيات العربيّة إلى خمسة تصنيفات هي: اسم(ism) وكنية(kunya) ونسب (nasab) ولقب(laqab) ونسبة(nisba). الاسم هو الذي يطلق على الطفل أو الطفلة عند الولادة مثل أحمد أو سميرة وغير ذلك، بينما الكنية صفة تشريفية تطلق على الرجل أو المرأة مثل عم وعمة وأب وأم. أما النسب فهو علاقة تربط بين المولود والوالد ويبدأ دائما ب»ابن أو بن و ابنه أو بنت» مثلا نقول "فهد بن عبد العزيز، ومريم بنت عمران"، أما اللقب فهو صفة تطلق على الشخص مثل الرشيد والصديق وغيرهما. وأخيرا النسبة تشير إلى الاسم الأخير للشخص أو اسم العائلة.

قدم هذا البحث نظاما لاستخراج أسماء الأعلام من النصوص المكتوبة باللغة العربيّة باستعمال المحول التتابعي، حيث اعتمد الباحثون على أداة كاسيس (CasSys) لبناء النموذج التتابعي، وهذه الأداة تمتاز بالتوافقية مع المنصة اللغوية العالمية يونيتكس (Unitex)، فقام الباحثون باستخدام أداة تسمى كيويكس (Kiwix) للحصول على الملفات على شكل نصوص من موسوعة الويكيبيديا. وفي نهاية المطاف اختُبِر النظام المقترح على المحتوى النصي المستخرج من هذه الموسوعة، وجرى تقييم النظام اعتمادا على مجموعة من مقاييس تقييم أداء الأنظمة المشهورة، وأثبتت النتائج كفاءة وقدرة النظام المقترح.

A. Ranta, B. Nordström (Eds.): GoTAL 2008, LNAI 5221, pp. 440–4512008 ,

Year of publication: 2008

Arabic Named Entity Recognition from Diverse Text Types

Khaled Shaalan, Hafsa Raza

التعرف على أسماء الأعلام العربيّة في أنواع نصوص مختلفة

في السنوات القليلة الماضية، اهتم الباحثون بموضوع التعرف الآلي على الأسماء في النصوص العربيّة سواء أسماء الأشخاص أم الأماكن أم غيرها، وأدرج هذا الموضوع في حقل معالجة اللغات الطبيعية. لكن عددا قليلا من هذه الأبحاث ركَّز على التعرف على أسماء الأعلام (NER) للنصوص العربيّة بسبب نقص الموارد للأنماط العربيّة وعدم الاهتمام الملموس في معالجة اللغة العربيّة الطبيعية بشكل عام. لذا قامت هذه الدراسة بتطوير نموذج لاستخراج الأسماء (الكيانات) للغة العربيّة (NERA)، حيث يستند هذا النموذج على قواعد يدوية تستخدم ثلاثة مكونات، وهي: القواميس، وقواعد اللغة (التعبيرات العادية)، بالإضافة إلى طرق آلية للترشيح.

اعتمدت عمليات الترشيح (filtering) على استبعاد الأسماء غير الحقيقية عن طريق رفض الكلمات من خلال قواعد معينة تتعلق بنهايات أو بدايات الجملة، كذلك رفض الأسماء ضمن قائمة أسماها الباحثون «القائمة السوداء». على سبيل المثال «رئيس القسم السابق مشرف العلوم» ينبغي عدم اعتبار «مشرف العلوم» على أنها اسم حسب نظام الترشيح المتبع في هذه الورقة البحثية. كذلك يلعب الترشيح الذي استخدم في هذا البحث دورا هاما في التحديد الصحيح لبعض أنواع الأسماء، فمثلا، شركة الفوسفات الأردنية، التصنيف الصحيح لها يكون على أنها اسم شركة (Company) وليست اسم موقع (Location).

ويدعم النظام المقترح ما مجموعه عشرة أنواع من الأسماء؛ وهي أسماء المواقع،

والأشخاص، والمنظمات، والأسعار، والتاريخ، والوقت، ورقم الهاتف، وأسماء الملفات، والقياسات. ولتقييم النموذج المقترح، قام الباحثون باستخدام بيانات تدريب مستخلصة من نصوص القرآن الكريم والمواقع الإلكترونية بالإضافة إلى الأدب العربي، وذلك من أجل بناء بيانات خاصة بهم، وأيضا لاستخراج المعلومات الدلالية بعمق ودقة. وبالنسبة لأسماء الأشخاص والمواقع والمنظمات، فقد حقق النموذج (٨٧,٧٪)، و(٨٥,٩٪)، و(٨٣,١٥٪) على التوالي. وقد أظهر النظام نسبة دقة تجاوزت (٩٠٪)، كمتوسط لجميع أنماط الأسماء، وذلك باستخدام نهج قائم على القواعد. ويتكون النظام من القائمة البيضاء التي تمثل قاموس الأسماء، والقواعد النحوية، في شكل التعبيرات العادية، التي تُعدُّ المسؤولة عن التعرف على أسماء الأعلام. وجرى تقييم نموذجهم المقترح الذي أطلق عليه اسم (NERA) باستخدام طريقة شبه آلية، باستخدام ذخيرة لغوية، وكانت نتائج الأداء مرضية من حيث الدقة، والاستدعاء، والمعدل التوافقي.

3rd International Conference on Arabic Language Processing (CITALA'09), May 4-5, 2009, Rabat, Morocco

Year of publication: 2009

Automatically Extending NE coverage of Arabic WordNet using Wikipedia

Musa Alkhalifa, Horacio Rodríguez

التوسعة الأوتوماتيكية لتغطية الكيانات المحددة لـ "ووردنيت" العربيّة باستخدام ويكيبيديا

يركز هذا البحث على الاستخلاص الأوتوماتيكي لتوسعة الكيانات المحددة (Named Entities (NE)) العربيّة باستخدام المعلومات المستخلصة من ويكيبيديا، وربطها أوتوماتيكيا بـ «ووردنيت» العربيّة والإنجليزية. تستند عملية جمع الكيانات المحددة في هذا البحث لتضمينها في «ووردنيت» العربيّة إلى طريقة شبه أوتوماتيكية، إذ تتألف من خطوتين: اختيار الكيانات المرشحة ثم التحقق منها يدويا.

في مرحلة اختيار الكيانات المرشحة يكون هدفنا هو تقييد عدد الكيانات الممكنة قدر الإمكان لتقليل الجهد البشري، وفي الخطوة الثانية، يركز البحث على معلومات من ثلاثة موارد هي قاعدة بيانات (GEONAMES) وهي قاعدة بيانات الأسماء الجغرافية لنأخذ منها المعلومات المتعلقة بالدول العربيّة، والفهرس الجغرافي لدول العالم من منظمة الأغذية والزراعة، ومدخلات الكيانات المحددة من معجم جامعة ولاية نيوميكسيكو، وبالطبع فإن هذه المدخلات فيها بعض الإشكاليات التي لا بد من معالجتها يدويا في الخطوة الثانية، وقد نتج عن هذه المرحلة ١،١٤٧ مجموعة مترادفات (synsets)، فيها ١،٦٥٩ متغير من ٣١ نوع.

عند وجود كيانات محددة إنجليزية فإنه يمكننا الحصول على مقابلاتها العربيّة باستخدام ويكيبيديا، إذ يمكننا استعادة الصفحة المقابلة للكيانات المحددة الإنجليزية إذا كانت موجودة

بالبحث عن الروابط التي تشير إلى مقالات بلغة أخرى في ويكيبيديا (interwiki link) وتحديدا تلك التي تشير إلى مقالات عربيّة، ولكن هناك بعض المشكلات التي ستواجهنا: أولا تحديد الكيانات المحددة الإنجليزية التي يجب أن نبحث عنها، فعند البحث مباشرة في ويكيبيديا فإننا سنحصل على الكثير من الصفحات غير المتعلقة بالكيانات المحددة، وعلينا توصيل (Map) الصفحة مع مجموعات مرادفات من "ووردنيت" الإنجليزية وحينها سنواجه مشكلة «فك اللبس لمغزى الكلمة» (Word Sense Disambiguation)، لذا نبدأ بـ "ووردنيت" الإنجليزية لنواجه مشكلة التوضيح مع ويكيبيديا وليس مع «ووردنيت»، ثانيا: كيفية التعامل مع تعدد الصفحات أي حين تكون عدة صفحات متصلة بكيان محدد، وهنا يساعدنا وجود صفحات توضيح في ويكيبيديا. ثالثا: الصفحات العربيّة على ويكيبيديا لا يوجد فيها تشكيل (unvowelized)، والمشكلة هنا أن «ووردنيت» العربيّة يجب أن يكون فيها تشكيل وذلك يمكن أن يُعمل بشكل يدوي ولكن الهدف هنا هو تقليل الجهد البشري.

تعمل هذه الطريقة أولا بجمع مجموعة اقتراحات «ووردنيت» الإنجليزية ثم وصلها بمقابلاتها، ثم نستخلص مجموعات المفردات المرشحة، وأهم ما في هذه الطريقة هو تصفية المجموعات المرشحة بالاستعانة بويكيبيديا الإنجليزية، ويجري ذلك لكل كيان محدد إنجليزي مع معلومات توضيحية بالبحث عنه في ويكيبيديا الإنجليزية، فإذا لم نعثر على صفحة فإنه ليس للكيان مقابل بالعربيّة، أما إذا عثرنا عليها فإننا نتبع تسلسل الروابط حتى نعثر على الصفحة الحقيقية، ثم نبحث عن رابط إلى صفحة عربيّة، وفي النهاية نقوم بإضافة تمييز أحرف العلة بطريقة أوتوماتيكية متحفظة تقسم الأمر إلى أربع حالات: في حالة الترجمة المباشرة من لغة أجنبية ويكون المصطلح العربي فيه حروف علة (ا ي و)، وفي حالة الترجمة أيضا ولكن تكون بعض حروف العلة القصيرة (حركات التشكيل)، وحالة تكون للكلمة لها مثيل في «ووردنيت» العربية، وحالة تكون فيها الكيانات المحددة العربيّة ليس لها مقابل في المصطلحات الأجنبية، وكل حالة لها حلها الخاص بها. وتمّ التحقق من دقة هذه الطريقة بمقارنتها بالطريقة اليدوية فجاءت النتائج مطابقة بنسبة ٩٣٪ وكان هناك خطأ بنسبة ١,٧٪ ، بينما كانت نسبة ٥٪ كانت

غير معروفة.

يقدم هذا البحث طريقة أوتوماتيكية لربط الكيانات المحددة العربيّة بالإنجليزية لـ"ووردنيت" العربيّة والإنجليزية ، باستخدام ويكيبيديا العربيّة والإنجليزية كموارد معرفة، فكان النظام أوتوماتيكيا بشكل كامل ودقيق تقريبا، وطُبق لإثراء مجموعة الكيانات المحددة في «ووردنيت» العربيّة.

Transactions of the Association for Computational Linguistics, vol. 3, pp. 243–255, 2015.
Year of publication: 2015

Combining Minimally-supervised Methods for Arabic Named Entity Recognition

Maha Althobaiti, Udo Kruschwitz, Massimo Poesio

الجمع بين الأساليب الخاضعة للحد الأدنى من الإشراف للتعرف على أسماء الأشياء بالعربيّة

يمكن للطرق الخاضعة للإشراف أن تحقق أداءً عالياً في مهام معالجة اللغة الطبيعية (NLP)، مثل تقنية التعرف على أسماء الأشياء ((Named Entity Recognition (NER) لكن هناك حاجة باستمرار إلى شروحات أو تعليقات (annotation) جديدة تلزم لكل تغيير يحصل في النطاق أو النوع. ولقد حفز هذا البحوث في مجال الطرق القابلة للإشراف الأدنى (Minimally supervised methods) كالتعلم شبه الإشرافي والتعليم عن بعد (distant learning). لكن لم تحقّق أي طريقة من هذه الطرق مستويات إنجاز مقاربة للطرق القابلة للإشراف حتى الآن.

قدمت الطرق شبه الإشرافية دقة عالية لكنها بطيئة الاسترجاع (low recall) بينما التّعلُّم عن بعد قدم استجابة سريعة لكن دقة أقل نسبياً. يشير هذا الاختلاف إلى أنه يمكن تحقيق نتائج أفضل من خلال دمج نوعين من الطرق الخاضعة للحد الأدنى من الإشراف.

في هذه الورقة قُدّمت طريقة مطولة للتعرف على أسماء الأشياء بالعربيّة (NER) من خلال مزج الطرق شبه الإشرافية وتقنيات التعلم عن بعد، كما جرت تجربة مصنِّف (Classifier) للتعرف على أسماء الأشياء بالطرق شبه الإشرافية، وتجربة آخرى باستخدام تقنيات التعلم عن بعد، ثم تدمج جميعها باستخدام مخططات مصنفة مركبة متنوعة (Classifier Combination Schemes) وتضم أيضاً إجراءات مزيج بايز للتصنيف(Bayesian Classifier Combiniation) (BCC) التي أعدت مؤخراً لتحليل الشعور (sentiment analysis).

ونتيجةً لذلك، فإن نموذج BCC قاد إلى زيادة الأداء بنسبة ٨٪ متفوقاً على أفضل مصنف

سابق .

الخلاصة:

أجريت أخيراً العديد من الدراسات في مجال حوسبة اللغة العربيّة فيما يتعلق بالتعرف على أسماء الأشياء، واستخدام النظم الخاضعة للإشراف واستكشاف مزايا مختلفة لاستخراج مجاميع معيارية لتصنيفات أسماء الأشياء تكون معنونة يدوياً (manual annotated standard corpora). تستغرق عملية استخراج أسماء الأشياء يدوياً وتصديرها إلى مجالات جديدة مزيداً من الوقت والجهد في الطرق الإشرافية الدنيا. لكن بإنجاز أقل من الطرق الإشرافية الأخرى.

تبين النقاط الآتية الخطوات العملية في هذه الدراسة:

- استعرضت طريقة طويلة لتقنية التعرف على أسماء الأشياء باستخدام مزيج بين التعلم شبه الإشرافي والتعلم عن بعد.
- استُخدم أيضاً مصنف (IBCC) للتعرف على أسماء الأشياء، وقورنت هذه الطريقة مع طرق التقييم التقليدية.
- أنتجت محددات توزيع المصنف (Classifier Combination restriction) للتحكم بكيفية التنبؤ للمصنف المتوقع للتشكيل وتوقيته.

أظهر هذا البحث استعمال تراكب بين الطرق الخاضعة للحد الأدنى للإشراف باستخدام طرق تمثيل مصنف لتقود إلى نتائج (NER). إن استخدام IBCC حسَّن من نسبة النتائج بزيادة ٨٪، بينما التحسين في الأداء باستخدام طرق التصويت (voting methods) هي فقط من ٤-٦٪.

وعلى الرغم من أن جميع نتائج طرق المزج (combination methods) صُنّفت بشكل دقيق، إلا أن نموذج IBCC حقق قابلية استرجاع بشكل أفضل من سائر طرق المزج بدون التأثير السلبي على الدقة. وبالإجمال فقد جرى تطوير نموذج مختلف وسهل للتعرف على أسماء الأشياء دون الحاجة إلى التدخل البشري (human intervention).

Proceedings Volume 10011, First International Workshop on Pattern Recognition; 100111I (2016); Tokyo, Japan, doi: 10.1117/12.2240887

Year of publication: 2016

Cross Domains Arabic Named Entity Recognition System

S. Saad Al-Ahmari, B. Abdullatif Al-Johar

أنظمة التعرف الآلي على أسماء الأشياء في العربيّة باستعمال تقاطع النطاقات

الملخص:

قام التعرف الآلي على أسماء الأشياء (named Entity recognition NER) بدورٍ هامٍ في العديد من تطبيقات معالجة اللغات الطبيعية (NLP) مثل: استخراج البيانات (Information Extraction IE)، وسؤال وجواب (Question Answering QA)، وتجميع النص (text clustering)، وتلخيص النص (text summarization)، وفك غموض كلمات الحس والشعور (word sense disambiguation).

هذه الورقة استعرضت تطوير نظام غير معتمد على النطاق (domain independent system) وتنفيذه للتعرف على ثلاثة أنواع لأسماء الأشياء في العربيّة. ويعمل النظام المعتمد على مجموعة قواعد لغوية مستقلة مع موسِّم أقسام الكلام (pos tagger) إضافة إلى قوائم الكلمات الثلاثية (trigger words) والصحافة. قام التعرف على أسماء الأشياء بدوراً هامٍ في العديد من تطبيقات معالجة اللغة الطبيعية وتمكنت طريقة التعرف على أسماء الأشياء من شمول مهمتين: الأولى إمكانية تحديد أسماء الأشياء بواسطة السياق، وهذا الأمر تضمن تعريفا صحيحا لبداية ونهاية الرموز (tokens) لكل اسم كائن مكون من عبارة (name entity phrase). الثانية تمكن من تصنيف الأسماء المعرفة إلى مجموعة طبقات معرفة مسبقاً (predefined classes)

كأسماء الأشخاص والأماكن.

الطريقة المعتمدة على القواعد، وهي إحدى الطرق المستخدمة للتحديد والتعرف على أسماء الأشياء، فهي تحتوي على مجموعة من القواعد النحوية والصرفية وقوائم معرفة مسبقاً كالصحف والكلمات الثلاثية. وتعتمد على شمولية الكلمات الثلاثية والقواعد التقنية التي يمكنها تحديد أسماء الأشياء وتصنيفها في مجالات محددة. في هذه الدراسة أنتجنا خمسة عشر مجالا مختلفاً للتعرف على أسماء الأشخاص، والأماكن والمنظمات، في النصوص العربيّة، والكلمات الثلاثية جُمعت من موارد ومجالات مختلفة.

هناك عدد كبير من الباحثين أنجزوا نظم تعرُّف على أسماء الأشياء بلغات مختلفة، إلا أن عددا قليلا من الباحثين ركزوا على البحث في اللغة العربيّة. التعقيد الصرفي للغة العربيّة جعل التعرف الآلي على أسماء الأشياء مهمة صعبة، فاللغة العربيّة تمتلك مزايا فريدة وخاصة بها لا توجد في اللغات الأخرى، كما أن هناك بعض المزايا المتشابهة والمتماثلة مع بعض اللغات السامية الأخرى.

الخلاصة والعمل المستقبلي:

استعرضنا في ورقتنا عملية تطوير نطاقات مستقلة في التعرف الآلي على أسماء الأشياء، والهدف الرئيسي من هذا العمل هو استخدام طريقة تعتمد على القواعد للتعرف على أسماء الأشياء بالإضافة إلى الأشخاص، والمواقع والمنظمات في النصوص العربيّة، وطريقتنا تحتوي على ثلاث خطوات رئيسة: المعالجة المسبقة – التحضيرية، وتوسيم أقسام الكلام (POS Tagging) وخوارزمية التعرف (recognition algorithm). وقد طُبّقت الطريقة على مجموعتين اثنتين في مجالين مختلفين من النصوص العربيّة، حيث استخدمت طريقة التجربة لحساب مؤشرات الأداء لكل ذخيرة (CORPUS) و أظهرت نتائج التجربة أن نتائج النظام أفضل بالمقارنة مع الأنظمة الأخرى، باستخدام طرق وأدوات الفحص، كما أثبتت أيضاً الاحتمالات المتوقعة لتقاطع المجالات اللغوية، كما أظهرت نتائج جيدة عند مقارنتها وفحصها بنطاقات أخرى.

وفي العمل المستقبلي نتطلع إلى توسيع استخدام الصحافة (gazatteers) وإضافة قوائم معرفة أخرى، لتحسين أداء النظام، كما أنّ هناك أيضاً احتمالية استخدام طرق ونظم تعلم آلي (machine learning) دون استخدام نظام يعتمد على القواعد وكيفية تأثيرها على نسبة الإنجاز الكلية للنظام.

(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 3, 2016

Year of publication: 2016

Integrating Semantic Features for Enhancing Arabic Named Entity Recognition

Hamzah A. Alsayadi, Abeer M. ElKorany

إضافة قواعد دلالية لتحسين أنظمة التعرف الآلي على أسماء الأشياء

الملخص

يُعدُّ التعرف الآلي على أسماء الأشياء (named entity recognition NER) حالياً حقلاً أساسياً للباحثين في مجال معالجة اللغة الطبيعية (NLP) ويهدف إلى إيجاد حل لتسريع وتحسين دقة تحديد أسماء الأشياء، وفي هذه الورقة استعرضنا نظام تعلم آلي دلالي (semantic machine learning system) لمشكلة التعرف على أسماء الأشياء في العربيّة.

الهدف الأساسي من عملية التعرف الآلي على أسماء الأشياء هو تحسين دقة إقرار أسماء الأشياء والتعرف عليها واستخراجها، وتُعدُّ هذه المهمة ضرورية في تطبيقات معالجة اللغة الطبيعية مثل: تجميع النتائج (results clusters)، والترجمة الآلية (machine translation)، ونظم التصفح (navigation system)، وتحسين استرجاع المعلومات (enhancing information retrieval)، وتحسين النتائج في أنظمة السؤال والجواب (question answering system).

نظراً لما تقوم به اللغة العربيّة من دور مهم وضروري على المستوى الاجتماعي والاقتصادي في السنوات العشر الأخيرة بدأ الباحثون في تطوير تطبيقات اللغة العربيّة مثل أنظمة التعرف الآلي على أسماء الأشياء في اللغة العربيّة، فهي لغة ثرية وغنية بالكلمات والنحو والتركيب، ولها تركيب غاية في التعقيد الصرفي. وهناك مستويات للغة العربيّة، وهي ثلاثة: اللغة العربيّة التقليدية، واللغة العربيّة المعيارية الحديثة، واللهجة العامية غير الرسمية (اللهجات).

تستخدم اللغة العربيّة الحديثة في هذه الدراسة، وهذا يتضمن العديد من التحديات كالغموض الصرفي، وتعقيد الأسماء المشتركة مع الكلمات وغموضها. الفكرة الأساسية لهذا النموذج هو دمج مزايا لغوية متعددة وتوظيف الدلالة في العلاقات الدلالية بين أسماء الأشياء، ويركز النموذج على ثلاثة حقول لتمييز الأسماء (الأشخاص، الأماكن، المباني). واعتمادا على ذلك ضمت مزايا داخلية (**internal features**)، مثلت القواعد اللغوية (**linguistic features**)، ومزايا خارجية (**external features**) التي يمكن أن تصف الدلالة بين العلاقات بين الأنواع الثلاثة لتحسين دقة التعرف على أسمائها باستخدام موارد معرفية خارجية مثل (**ANW Arabic wordnet ontology**).

أظهرت النتائج العملية أن نسبة الإنجاز الكلية للنظام من الممكن أن تصل إلى ٨٦ , ٨٧٪، و٧٢ , ٨٤٪ لكل من مجموعتي **ANERCORP** و **ALTEC**.

الخلاصة:

هذه الورقة استعرضت دمج وتكامل مجموعة المزايا لغرض التعرف على أسماء الأشياء، وهذا التكامل دمج المزايا الداخلية التي مثلت المزايا اللغوية واعتبرتها مزايا خارجية لتمثيل الدلالة في اللغة العربيّة. والمزايا الداخلية هي مثل **POS**،**GAZ** والمؤشر (**indicator**) ومزايا حروف **Cram**، بينما المزايا الخارجية هي مزايا معلومات دلالية استخرجت من نظام **Arabic wordnet ontology** مثل الطبقات (**classes**) والمباشرة (**instance**) والعلاقات (**relations**).

النموذج المتكامل ساعد في حل بعض التعقيدات اللغوية والصرفية والإملائية في اللغة العربيّة، ونتائج التجارب بينت أن **F-measure** لـ **ANERCorp** و **ALTEC** حوالى (٨٦ , ٨٧٪ و ٧٢ , ٨٤٪) ومجموعة المزايا المفترضة حققت تحسناً على أبحاث سابقة بنسبة ٥٦ , ٣٪ في التعرف على أسماء الأشياء. في المستقبل، نتطلع إلى دراسة إمكانية تحسين أداء النظام باستخدام طرق أخرى مثل نظم معتمدة على القواعد ونظم هجينة مع مزايا المعلومات الدلالية.

Proceedings of Recent Advances in Natural Language Processing, pages 731–738, Hissar, Bulgaria, Sep 7–9 2015.

Year of publication: 2015

Named Entity Recognition of Persons' Names in Arabic Tweets

Omnia H. Zayed, Samhaa R. El-Beltagy

التعرف على أسماء الأشياء في أسماء الأشخاص في التغريدات العربيّة

الملخص:

إن انتشار استخدام اللغة العربيّة في وسائل التواصل الاجتماعي لا سيما في تويتر (Tweeter)، أدى إلى اهتمام متزايد في بناء تطبيقات معالجة اللغة الطبيعية للغة العربيّة وخاصة بما يتناسب مع اللهجات العامية (colloquial)، وهو نموذج شائع الاستخدام في وسائل التواصل الاجتماعي. إن الخصائص الفريدة للغة العربيّة تجعل من عملية استخراج أسماء الأشياء تحديا كبيرا، حيث إن طبيعة التغريدات في برنامج تويتر تضفي أبعاداً جديدة.

كما أن شهرة البحوث المنجزة سابقاً في التعرف الآلي على أسماء الأشياء ركزت على استخراج الأشياء من اللغة القياسية المسماة (اللغة العربيّة المعيارية الحديثة) (modernstandard Arabic MSA)، كما أن طبيعة اللهجات العامية المستخدمة في التغريدات تقلل من كفاءة نظام التعرف الآلي على أسماء الأشياء المطورة أصلاً للتعامل مع النصوص العربيّة المعيارية الحديثة.

ركزت هذه الورقة النظر على مهمة التعرف الآلي على أسماء الأشخاص (Person's name recognition) من التغريدات دون إجراء أي تحليل صرفي (morphological analysis) أو مزايا تعتمد على اللغة نفسها (language depends features). كما أن النظام المفترض تبنى نموذجاً مبنياً على القواعد (rule-based model) مدمجاً مع نموذج إحصائي

آخر (statistical-based model).

هذه الطريقة استخدمت طريقة غير خاضعة للإشراف (unsupervised) لتعلم الأنماط (patterns) كما استعملت المعاجم العنقودية (Clustered dictionaries) للتعرف على أسماء الأشياء وحل غموضها. كما تفوقت الطريقة المعروضة على أفضل النتائج المسجلة في المراجع لنفس الفحص فقد سجلت نسبة زيادة مقدارها ٦ , ١٩٪ في (F-score).

الخلاصة والعمل المستقبلي:

قدمت هذه الورقة طريقة لاستخراج أسماء الأشخاص وتحليل غموضها، وكان الاهتمام الأكبر أثناء عملية البناء والتطوير هو محاولة حل الغموض المتوارث لأسماء الأشخاص العربيّة دون استخدام موارد تعتمد على اللغة أو بالاعتماد على موارد لغوية واسعة (extensive lexical resources).

الهدف الرئيسي هو محاولة شمول النظام للمجالات واللغات وأنواع النصوص الأخرى، وهذه الطريقة جمعت بين الأسماء من المعاجم وعناقيد الأسماء (name clusters) مع نموذج إحصائي لاستخراج نماذج سياقية (unigram) من خلال طرق غير خاضعة للإشراف، التي استخدمت أساساً في اختيار أسماء الأشخاص. والفكرة الرئيسية التي اتبعت في هذه الدراسة هي تعليم قيود توافقية من خلال عنقدة (clustering) الأسماء والأنماط المسجلة (scored pattern) وقد جرى استغلال قائمة الأسماء الكاملة التي جمعت من خلال موارد عامة متوفرة، وقُيِّمَتْ الطريقة المعروضة جُرِّبتْ فظهر أنها تفوقت على جميع المحاولات الحالية لاستخراج أسماء الأشياء باللغة العربيّة من التغريدات. في المستقبل نخطط لتوسعة هذا المشروع لاستخراج أسماء أشياء أخرى كالمواقع والمؤسسات والمنظمات.

Natural Language Engineering, 23(3), 441-472. doi:10.1017/S1351324916000097

Year of publication: 2017

NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic

Mai Oudah, Khaled Shaalan

تحسين التغطية والأداء لنظام التعرف على أسماء الأشياء بالاعتماد على القواعد في اللغة العربيّة

تُعدُّ مهمة التعرف الآلي على أسماء الأشياء (Named Entity recognition) مهمة أساسية في العديد من أنظمة معالجة اللغة الطبيعية، وتعزز استخدام الموارد اللغوية المختلفة. وتصبح مهمة التعرف الآلي على أسماء الأشياء مهمة معقدة وصعبة جداً للغات التي تملك بنية معقدة وهيكلية صعبة، إضافة إلى ثراء صرفي يزيد الأمر تعقيداً كما هي اللغة العربيّة، فهذه اللغة تمتلك العديد من الخصائص التي تجعلها في قمة التحدي.

في أعمال سابقة افترضنا نظام تعرف آلي يتخذ منهجا هجيناً كدمج كل من الطريقة المعتمدة على التعلم الآلي (machine learning approach) مع الطريقة المعتمدة على القواعد(rule-based approach). وفي هذه الورقة قدمنا طريقة مطورة جديدة لزيادة شمول الأنظمة المعتمدة على القواعد للتعرف على أسماء الأشياء للسعي في تطوير أدائها وإنجازها والسماح لها في تحديث القواعد المستخدمة تلقائياً.

الآلية المستعرضة استخدمت قرارات التعرف الناتجة عن النظام المهجن لتحديد ضعف الطريقة المعتمدة على القواعد واشتقاق قواعد لغوية جديدة تهدف إلى تحسين أساس القاعدة التي من الممكن أن تحقق مرونة أكبر ودقة نتائج أفضل. كما استخدمت مجموعة ACE المعيارية كمورد لاستخراج وتحليل قواعد لغوية جديدة للتعرف على أسماء الأشخاص والمباني

والأماكن، واشتُق أربعة عشر نمطاً جديداً، صيغت على شكل قواعد لغوية وجُرّبت وقُيّمت من حيث التغطية والشمول. النتائج التجريبية أظهرت أن أداء النظام المحسن المبني على القواعد أنجز تغطية أسماء الأشخاص والأماكن والمباني المفقودة حسب الآتي على التوالي: ٩٣,٦٩٪و٠٩,٥٧٪و٢٨,٥٤٪.

الخلاصة:

في هذه الدراسة قدمنا طريقة جديدة لاستخراج القواعد اللغوية للتعرف الآلي على أسماء الأشياء بهدف تحسين أداء المكون المعتمد على القواعد في النظام المهجن للتعرف على أسماء الأشياء. والتقنية الآلية لصيانة النظام المبني على القواعد سيساعد على حذف وإخفاء العيب الرئيسي في النظام المبني على القواعد. ومع تحسين أداء النظام المبني على القواعد سيكون هناك نتائج أدق وأفضل. نظامنا المهجن قدم مكونين اثنين، أحدهما مبني على القواعد والآخر مبني على التعلم الآلي، وهو مؤهل للتعرف على أحد عشر نوعا مختلفا من أسماء الأشياء .

الطريقة المفترضة تعتمد على ثلاث خطوات رئيسية: استخراج عوامل المزايا لأسماء الأشياء التي فشل نظام NERA في تحديدها، لكن نجح النظام الهجين في تصنيفها. الثانية تحليل مخرجات الخطوة السابقة لكشف الأنماط المرشحة للتعرف الآلي على أسماء الأشياء. والخطوة الثالثة تقييم أداء الأنماط المرشحة لإيجاد الأنسب منها.

مجموعة بيانات ACE تستخدم بعدها مجموعة بيانات مصدرية لاستخراج واشتقاق القواعد اللغوية للتعرف على أسماء الأشخاص والمواقع والمباني حسب دلالتها بصفتها المجموعة الكبرى للمعلومات المعيارية المستخدمة في أنظمة التعرف الآلية على أسماء الأشياء في مجتمع الدراسة.

The International Arab Journal of Information Technology, Vol. 6, No. 5, November 2009

Year of publication: 2009

Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition

Yassine Benajiba, Mona Diab, Paolo Rosso

دعم التعرف على أسماء الأشياء باستعمال مؤشرات لا تعتمد على اللغة ومؤشرات خاصة

حظيت مهمة التعرف الآلي على أسماء الأشياء باهتمام كبير؛ لأنها قدمت تحسينات كبيرة وبارزة في العديد من تطبيقات معالجة اللغات الطبيعية، ونحن على وشك مشاهدة تقدم كبير في تطوير أنظمة التعرف على أسماء الأشياء للغات غير الإنجليزية، لوفرة الموارد العربيّة نسبياً من جهة والنضوج الكبير في معالجة اللغة العربيّة، فمن الطبيعي مشاهدة اهتمام في تطوير تقنية التعرف الآلي على أسماء الأشياء للغات بشكل عام. في هذه الورقة جرت دراسة تأثير مميز لاستخدام مجموعات مختلفة من الخصائص اللغوية للغة مستقلة ومتخصصة في إطار تعلم الآلة، وهي آلات دعم القطاع (support vector machines).

لقد اكتُشفت مزايا لغوية وصرفية وسياقية (دلالية) وتسع مجموعات بيانات لأنواع وتعليقات مختلفة، حيث قسنا حاسوبياً تأثير المزايا المختلفة منفصلة أو مشتركة، فسجلت أعلى نسبة إنجاز من خلال دمج جميع الخصائص FI=٧١,٨٢. وبشكل أساسي فإن دمج الخصائص اللغوية المستقلة مع خصائص اللغة يقود إلى نتائج أفضل في جميع الأنواع في النصوص التي جرى فحصها ودراستها. ومع ذلك نلاحظ على مستوى الطبقة (class level) أن الطبقات المختلفة لأسماء الأشياء تتأثر بشكل مختلف من المزايا الصرفية المطبقة.

الملخص والتوجهات المستقبلية:

وصفنا نظام تعرف آلي على أسماء الأشياء باستخدام آلات دعم القطاع (SVMs) مع دمج لكل من استقلالية اللغة والمزايا اللغوية المستقلة للتعرف على أسماء الأشياء بالعربيّة، فقد قمنا بقياس تأثير المزايا المستقلة المختلفة في مجموعة مشتركة مرتبطة من خلال مجموعات بيانات معيارية مختلفة وأنواع مختلفة. وقد تفوقت تجاربنا على بيانات خط القاعدة (baseline) بشكل ملحوظ، فقد سجلت أفضل نتائجنا (١٧ , ٨٢٪) باستخدام جميع المزايا في بيانات ACE ٢٠٠٣ BN، وقد افترضت هذه النتائج أن الخصائص اللغوية المحددة المستخدمة بالاشتراك مع اللغة المستقلة مفيدة جداً في أنظمة التعرف على أسماء الأشياء. وقد أظهرت النتائج علاقة المزايا الصرفية للغات التي تتميز بالتعقيد والبنية الغنية.

علاوة على ذلك، فإن المزايا التي استخدمت في الدراسة جميعها كانت غير معتمدة على اللغة (language independent) باستثناء تلك التي استخرجت من خلال محلل صرفي وأداة MADA لفك الغموض، وسمحت باستخراج ١٤ مزية صرفية لكل كلمة. وظهر أيضا تحسن عند استخدام MADA في النصوص التي تظهر فيها أسماء الأشياء في سياقات مختلفة، وعموماً فإن استخدام أداة MADA قد سجل تحسيناً بنسبة ٨ , ٠٪ في المعدل على باقي المعجم فقط، وبعض هذه المزايا كالرقم، والجنس كانت مفيدة جداً لاستخراج أسماء الأشياء الطويلة لأنها تشير إلى نموذج مكوّن من سلسلة طويلة من الكلمات تشترك في نفس قيمة هذه المزايا، وعلى صعيد آخر، فهي أيضاً مفيدة في استخراج أسماء الأشياء التي لم تظهر في المجموعة التدريبية والكلمات الغريبة في اللغة العربيّة، ومعظم هذه الكلمات لا يمكن استخراجها من خلال المحلل الصرفي MADA.

4th International Conference on Arabic Language Processing, May 2–3, 2012, Rabat, Morocco

Year of publication: 2012

Using Wikipedia as a Resource for Arabic Named Entity Recognition

Fahd Alotaibi, Mark Lee

استخدام الويكيبيديا موردًا للتعرف إلى أسماء الأعلام العربيّة

يُعدُّ التعرف الآلي إلى أسماء الأعلام من الحقول المهمة في مجال الكشف عن الأنماط (الأعلام) الواردة في السياق وتصنيفها من الناحية الدلالية، سواء دلت على شخص أم منظمة أم اسم مكان أم غير ذلك. غير أن تعيين اسم النمط ليس مهمة واضحة حتى بالنسبة للغات التي لديها أدلة ومؤشرات تحدد نوع الكلمة، مثل الكتابة بالأحرف الكبيرة في اللغة الإنجليزية التي تدل على أسماء الأعلام وغيرها، ومع ذلك فإن بعض الأسماء تتطلب جهداً كبيراً لتمييزها مثل أسماء الكتب والأفلام أو العناوين الكبيرة. ولكن تصبح هذه المشكلة أكثر صعوبة في اللغات التي لا تحتوي على دلائل كتابية مثل اللغة العربيّة. وبشكل عام لا يوجد في اللغة العربيّة مؤشر يساعدنا على التمييز بين أسماء الأعلام والأسماء العامة أو حتى الأفعال، مما دفع الباحثين للتنقيب في موارد المعلومات المختلفة المحلية والعالمية لحل هذه المشكلة. وعلى الرغم من أن هناك عددا من المحاولات لمعالجة التعرف على أسماء الأعلام العربيّة، ما زالت جميع المحاولات تقتصر على وكالات الأنباء وبعض الموارد المتوفرة ، لذلك أصبح الدافع لدى الباحثين على هذا العمل هو توسيع كل من قدرة وآفاق اللغة العربيّة في هذا المجال خارج نطاق وكالات الأنباء. وقد استُخدمت الويكيبيديا مرجعا للعمل عليها في هذا البحث بسبب ثراء وتنوع المواضيع فيها.

في هذه الورقة البحثية قام الباحثون بوصف نهج جديد لتحديد أسماء الأعلام في الويكيبيديا

العربيّة. ولتحقيق ذلك قُدم تحليل لغوي لهذه الأنماط وتوبع تكرارها في الويكيبيديا العربيّة من حيث التغطية والتعقيد. وقد استخدموا خوارزمية تعلم الآلة كأداة تصنيف للكلمات للتنبؤ بوجود هذه الأنماط في الويكيبيديا العربيّة من خلال النظر إلى السمات النحوية، والدلالية، والسياقية والصرفية، حيث تقوم أداة التصنيف بعد تطويرها بالكشف عن حدود العبارات وتحديد الكلمات التي حدث تكرارها في النص. وقد أجريت مجموعة من التقييمات الإحصائية في شروط التردد لدراسة توزيع هذه العبارات وفقاً لمستوى تعقيدها. علاوة على ذلك، قُيّم التعرف على هذه الأعلام والعبارات في الويكيبيديا وطريقة توزيعها.

استفاد الباحثون من المخطط الأساسي للويكيبيديا عن طريق إيجاد الروابط المشتركة، حيث قاموا بتعيين الروابط المتداخلة علامةً إيجابيةً لتعيين الأنماط. وقد ثبت أن هذا النهج ذو فاعلية كبيرة خاصة في العبارات المعقدة. وقُيِّمت الطريقة الموصوفة على عينات عشوائية من حيث البساطة والتعقيد من النصوص في الويكيبيديا العربيّة، وحققت طريقتهم نسبة نجاح جيدة في التعرف على أسماء الأنماط (٦٢٪, ٨٨).

٣-٤-٤ أبحاث التشكيل الآلي

وتضم ثمانية أبحاث، منها خمسة أبحاث نوع (أ) هي: تحسين التّشكيل في اللغة العربية من خلال تحليل قواعد النّحو، وتشكيل اللغة العربية باستعمال مزيج موزون من الكلام مع النماذج المستندة على النص، وتشكيل اللغة العربية في سياق الترجمة الآلية الإحصائية، والتشكيل الآلي للغة العربية باستخدام الشبكات العصبية المتكررة، واستعادة آلية/ تلقائية لعلامات التشكيل العربية: منهجية بسيطة وإحصائية بحتة.

وثلاثة أبحاث نوع (ب) هي: طريقة هجينة لبناء مشكّل عربي، ونظام هجين للتشكيل العربي الآلي، والاسترجاع التلقائي للتشكيل في اللغة العربية اعتمادا على الشبكات العميقة.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1309–1315, Lisbon, Portugal, 17-21 September 2015.

Year of publication: 2015

Improving Arabic Diacritization through Syntactic Analysis

Anas Shahrour, Salam Khalifa and Nizar Habash

تحسين التّشكيل في اللغة العربيّة من خلال تحليل قواعد النّحو

المقدّمة

في العادة، تُغفِل اللغة العربيّة القياسية الحديثة علامات التشكيل التي تستخدم الترميز لمعرفة المفردات والتركيب النحوي. وتكمن مهمّة التشكيل الآلي للغة العربيّة في الاستعادة التلقائيّة للتشكيل المفقود، حيث ينطوي تحسين التشكيل في اللغة العربيّة على انعكاسات مهمة على المعالجة النهائية للغة العربيّة الطبيعية، على سبيل المثال، إدراك معنى الكلام وتركيب الكلام والتّرجمة الآليّة.

في السابق كانت الجهود مُنصبّة على استخدام الأساليب أو التقنيات الصّرفيّة في إزالة غموض صيغ المفردات، لكن لوحظ أنه رغم أن هذه التقنيات تعمل بشكل جيّد نسبياً على تشكيل المفردات، إلاّ أنّها تُصبح أسوأ بكثير بالنّسبة لتشكيل الحالة النّحوية (عادة على نهاية الكلمة أو المفردة)، ورأى الباحثون أن التحليل النّحوي قد يستفيد من التّشكيل الآلي، لكن يجب أن يقتصر الأمر على اختبار الفكرة، وبدلاً من ذلك فهو يُثبت أن هناك حاجة إلى سمات وقواعد لُغوية مُعقّدة لنموذج الحالة الإعرابية المعقدة باستخدام تحليلات نحوية تكون أكثر عُمقاً.

ناقشت هذه الورقة كيفية تحسين جودة التّشكيل الآلي في اللغة العربيّة بواسطة استخدام

أسلوب تحليل المفردات التّلقائي، الذي يجمع بين قواعد الكتابة لكل مقطع ومدى تطابقه مع التعلّم الآلي للحالة وضبطها على أساس قواعد علامات الصّرف.

مشكلة البحث والمبررات

تُعدُّ المعالجة الآليّة للغة العربيّة – خاصةً التّشكيل – أمراً صعباً، لأسباب عدّة:

أولاً: لأنّ اللغة العربيّة غنيّة بأشكال متعدّدة من حالات الصّرف، حيث يُقدّم المحلل الصّرفي الذي استُخدِم في هذه الدراسة مفردات عربية ذات ١٥ سمة، وتركّز هذه الورقة على سمتي الحالة والهيئة.

ضمن مجموعة البيانات التي استُخدمت في هذه الدراسة، أُعطيت سمة الحالة خمس قيم: حالة الرّفع (n)، وحالة النّصب (a)، وحالة الجرّ (g)، وحالة المطلق (غير المحدّد) (u) وأخيراً حالة غير المستعمل أو غير القابل للتطبيق (na). ويُعبّر عن حالات الرفع والنّصب والجرّ باستخدام مقطع صرفي (علامة ضبط) واضح. بينما استخدمت حالة المطلق (غير المحدّد) (u) للدّلالة على الكلمات التي ليس لها مقطع صرفي مُحدّد يُعبّر بشكل واضح عن حالة المفردة أو تلك التي ليس لها شروح مُخصّصة في الحاشية، حيث تكون معظم التعيينات المفقودة للأسماء الأجنبية التي غالباً لا تكون لها علامات حالة مُخصّصة. أما حالة غير المستعمل أو غير القابل للتطبيق (na) فقد استخدمت لغير الأسماء.

أما الهيئة فهي خاصيّة اسميّة ذات أربع قيم: المحدّد (d)، وغير المحدّد (i)، والمبني (c) وأخيراً هيئة غير المستعمل أو غير القابل للتطبيق (na). وتعكس الهيئة بشكل عام الوضوح أو الإتقان في الأسماء (d) مقابل (i) وتُبيّن فيما إذا كان الاسم هو عين الإضافة. في حين تُعبّر الهيئة (na) عن غير الأسماء.

ثانياً: تُعدُّ المفردات العربيّة غير المشكّلة غامضة للغاية؛ ففي البيانات التي استُخدمت في هذه الورقة، كان متوسط عدد التحليلات المحتملة للكلمات العربية (٨ , ١٢) تحليلاً لكل كلمة، ومعظمها مرتبط بحالات مُختلفة من التّشكيل. وتعكس بعض فروقات التشكيل

وجود أساليب مختلفة من فرضيات التّشكيل، في حين يُعزى البعض الآخر لاختلاف الصّرف والإعراب.

ثالثاً: تتضمن اللغة العربيّة حالات/ وهيئات مُعقّدة من التّخصيصات وأنماط التوافق التي تتفاعل مع إعراب الكلمة. فعلى سبيل المثال، يمكن للاسم أن يحصل على الحالة من خلال كونه فاعلاً لفعل ما، وأن يحصل على الهيئة من خلال كونه مُضافاً. في حين أنّ الصفات التي تُقيّد معنى الاسم تتفق معه في الحالة، تتحدّد هيئتها من خلال العنصر الأخير في سلسلة الإضافة التي يتقدّمها الاسم.

الاستعراض المرجعي

أجريت العديد من الدراسات السابقة على موضوع التّشكيل في اللغة العربيّة، حيث قام العديد من الباحثين بدراسة الموضوع. واستخدمت معظم تلك الدراسات تقنيات نمذجة تسلسلية (**sequence modeling techniques**) مختلفة تعتمد على درجات متفاوتة من المعرفة بدءاً بالحروف السطحية (**shallow letters**) وأشكال أو صيغ الكلمة وانتهاءً بالمعلومات الصّرفية الأعمق. واعتبر أحدُ الموارد التّشكيلَ جُزءاً من مشكلة إزالة غموض الصّرف (**morphological disambiguation problem**)). وعمد الباحثان في ذلك المورد إلى اختيار علامة الإعراب المثلى للمُفردة العربيّة ضمن السّياق، واستخدامها للاختيار من ضمن التّحاليل الممكنة للمحلّل الصّرفي. وقد اعتمدت هذه الورقة البحثية بشكل كبير على دراستين سابقتين وزادت على ما قامتا به من عمل من خلال استخدام سمات نحوية إضافية لتحسين دقة إزالة الغُموض الصّرفي. كما أضافت الورقة مزيداً من التحسينات من حيث خيار التحليل الصّرفي التّام (استخدام الفرضيات المساعدة، والتّرميز، وجميع السمات) بالإضافة إلى تشكيل الكلمة.

منهجية البحث

اعتمد هذا البحث على دافع تحليل الخطأ الذي جُرِّبت على ١٢٠٠ كلمة من مُخرجات

نظام **MADAMIRA**. وعثر الباحثون على عدد كبير من الأخطاء النّحوية المحتملة المفاجئة، مثل الأسماء المجرورة بعد الأفعال أو الأسماء المبنيّة المتبوعة بأسماء غير مجرورة أو غير مُضافة. قام الباحثون بتفسير هذه الأخطاء عن طريق نموذج التحليل السياقي (**contextual model**) المذكور، الذي يقتصر على نافذة صغيرة من الكلمات المجاورة بدون نمذجة بناء الجُمل، الأمر الذي أدّى إلى أداء أسوأ بكثير على الحالة وعلى التّشكيل مقارنة بأسلوب نظرية الفرضية المساعدة (**lemmas**) ونظام **POS** . (أقل بعشر نقاط مئوية انخفاضاً من ٩٦٪ إلى ٨٦٪).

اعتمد أسلوب هذا البحث على توفير تنبُّؤ أفضل حول الحالة والهيئة باستخدام نماذج الوصول إلى معلومات إضافية، لا سيّما التحليل النّحوي والقواعد. ومن ثمّ استخدام قيم الحالات والهيئات المُتوقعة لإعادة وضع علامات على مُخرجات نظام (**MADAMIRA**) عن طريق اختيار أفضل خيار من بين التحليلات الصّرفية المُصنّفة إلى رُتب.

النتائج والتوصيات

استعرضت هذه الورقة نتائجها التّجريبية من خلال المقارنة بين خمس تقنيات للتنبؤ بالحالة- الهيئة، حيث قورنت نتائج هذه التقنيات مع نظام خط الأساس للهيئة، مُقارنة بعدد من الأساليب الأخرى. واعتمدت الورقة على تقنيات التّصنيف الآتية:

قواعد الصّرف: أنشأ الباحثون دليلاً بسيطاً للتصنيف المعتمد على صرف- الكلمة، للتعامل مع أبرز أخطاء الحالة في الدراسة التجريبية.

المُصنّف المعتمد على قواعد الصرف: جُرب مصنّف للتنبؤ بتصحيح خط أساس حالة الكلمة وهيئتها باستخدام قائمة بيانات **DevTrain**. واستُخدمت فيه السمات نفسها التي اعتمد عليها المصنّف المعتمد على الصّرف.

قواعد الإعراب: وقد أدّت هذه القواعد إلى تحسين مستوى دقّة خطّ الأساس بنسبة ٥ , ٤ بالمئة للأسماء، بينما لم تُعط أيّ نتائج تذكر بالنسبة لمجمل الكلمات.

المُصنّف المعتمد على الإعراب: استخدم المصنّف للتنبؤ بتصحيح حالة الكلمة وهيئتها

ضمن قائمة بيانات **DevTrain** المعربة.

توليفة من قواعد الإعراب والمُصنّف المعتمد على الإعراب:

خلُصت نتائج الدراسة بعد استخدام أسلوب الاختبار العلمي إلى قُدرة النظام على زيادة دقّة تشكيل مجمل الكلمات بنسبة ٥ , ٢ بالمئة، في حين أنه ساعد على زيادة دقّة تشكيل الأسماء بنسبة ٢ , ٥ بالمئة، كما أنّ الانخفاض في نسبة الخطأ مقارنة مع برنامج (أوراكل) الشهير كانت ٣٠٪ و ٣٤٪ على التوالي.

وأوصت الدراسة في الختام أن يجري في المستقبل التحقيق في تطوير التشوّه الصرفي ونماذج التحليل النحوية. والعمل أيضاً على تحسين جودة تحليل الإعراب العربي الذي يقف وراء العديد من الأخطاء وفقاً لأخطاء التحليل التي أظهرتها الدراسة. وتشمل الاتجاهات المحتملة الأخرى استخدام تقنيات أكثر تطوراً للتعلم الآلي وسمات معجمية أكثر ثراءً.

INTERSPEECH'2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, 2334-2337

Year of publication: 2012

A Weighted Combination of Speech with Text-based Models for Arabic Diacritization

Aisha S. Azim, Xiaoxuan Wang, Khe Chai Sim

تشكيل اللغة العربية باستعمال مزيج موزون من الكلام مع النماذج المستندة على النص

وظّفت معظم دراسات التشكيل العربي خصائص الاستدلال النصي (**textually features inferred**) فقط. هذه الورقة تقترح طريقة جديدة، فتستخدم مزيجا موزونا من الكلام مع نموذج مستند على النص (**text-based model**) لتتيح للمعلومات الصوتية غير الحسّاسة لغويًا بأن تصحح وتكمل الأخطاء الناتجة عن نماذج توقّع التشكيل للنصوص (**diacritic predictions**). يعتمد النموذج الصوتي على نماذج ماركوف المخفية (**Hidden Markov Models**) ويعتمد النموذج النصي على الحقول العشوائية المشروطة (**Conditional Random Fields**) وهذا المزيج يؤدي إلى تناقص كبير في نسبة الخطأ خصوصًا في العلامات الإعرابيّة وهي الأصعب في معرفتها.

استُخدم محلل باكولتر الصرفي للغة العربيّة (**Buckwalter Arabic Morphological Analyzer(BAMA)**) بشكل كبير في الصرف العربي، لأن بإمكانه توليد قائمة من التصاريف المتوقعة لكلمة ما، وبالتالي يكون التنبؤ بالتصريف الصحيح هو عملية اختيار من قائمة.

يقبل النظام في هذه الورقة نوعين من البيانات: نص باللغة العربيّة، وإشارات صوتية تنطق النص بطريقة صحيحة، وسنوظّف مشكّلين آليين (**diacritizers**): الأول معتمد على

النص وممثل بحقول شرطية عشوائيّة (Conditional Random Fields)، والثاني معتمد على النطق وممثل بنماذج ماركوف المخفية (Hidden Markov Models (HMM)).

تتم العملية كالتالي: عند إدخال كلمة فإنّ BAMA سيولّد مجموعة من الحلول، ونحن هنا معنيون بإيجاد تسلسل مميز من التشكيل لسلسة من الأحرف الساكنة، وليس بالتحليل الصرفي، وبالتالي إذا أنتجت BAMA عدة احتمالات صرفية للكلمة، وجميعها بنفس تسلسل التشكيل، فإننا نعتبر أنّ الكلمة لها حل واحد. ومن ثم نأخذ الحل أو الحلول لتكون مدخلات للمشكّلين الآليين، ونستخدم النموذجين الصوتي والنصّي لتوليد نتيجة لكل حل، مما ينتج مجموعة من الحقول (tuples) لكل كلمة.

وصف النظام:

المشكّل الآلي المعتمد على النطق: اختُبر النموذج الصوتي على ما يقارب ساعتين ونصف الساعة من الكلام المسجّل من متحدث واحد في بيئة خالية من الضجيج، وباستخدام التشكيل الصحيح مع استشارة عالم في اللغة العربيّة، وقد احتوى النموذج على جميع أحرف اللغة العربيّة.

ثم يبدأ عمل نموذج التشكيل المعتمد على النطق بأربع خطوات:

تجميع احتمالات التشكيل جميعها باختيار جميع الحلول المميزة من مخرجات BAMA.

تحويل الحلول إلى نسخ صوتية باستخدام (G2P) القائم على القواعد (rule-based grapheme-to-phoneme) لتحويل كل حل ممكن إلى ملف صوتي ملائم.

تنفيذ عملية محاذاة لاستخلاص حدود الكلمة؛ فنستخدم نموذجًا صوتيا ونطبق محاذاة بين مدخلات الكلام والنسخ الصوتية للبيانات التي كانت نتيجة لتطبيق الخطوة السابقة.

تقييم نتيجة الاحتماليّة، وذلك بتنفيذ محاذاة (force-alignment) أخرى بين الإشارة الصوتية التي تعبر عن الكلمات غير المشكّلة والنسخ الصوتية لكل حل، وهنا يعطي مميز الكلام نتيجة احتمالية السجل الصوتي لكل نسخة صوتية لكل حل باستخدام مصفوفة فايتيربي (Viterbi)، ويُختار الحل عن طريق مشكّل آلي معتمد على الكلام، ويكون الحل هو صاحب

أعلى احتمالية في السجل.

المشكّل الآلي المعتمد على النص (diacritizer based-Text)

بُني نموذج النص باستخدام (Conditional Random Fields CRFs) التي تعد قوية في استخلاص السياق، لأنها تبسط افتراضات الاستقلالية المتأصّلة فيHMMs، وبالتالي قادرة على تحديد الاحتمالات المشروطة بين الملصقات والملاحظات، ويجري تطبيع (normalize) النص بحيث يُجرى تسقيط (mapping) على شكل واحد لواحد بين الملاحظات والملصقات، وذلك باستخدام الخصائص التالية: نافذة تدريب مكوّنة من تسعة أحرف ساكنة وهي الحرف الحالي وأربعة إلى يمينه وأربعة إلى يساره، والكلمة الحالية والسابقة والتالية، وعلامات تجزئة الكلام (POS) للكلمة الحالية والسابقة والتالية، كما يحتوي النموذج على أزواج (bigram) من أحرف العلة. وقد اختُبر نموذج النص على ما يقارب ٤٧٠ ألف كلمة مأخوذة من مقالات صحيفة النهار.

في التشكيل المعتمد على النص كانت الكلمة المدخلة عبارة عن سلسلة من الأحرف العربيّة الساكنة، وكل حرف يمكن أن نعيّن له واحدا من ١٥ ملصقا محتملا، وهذه المجموعة من الملصقات تتألف من جميع التشكيلات الممكنة للحرف، ويمكن احتساب الاحتماليّة الهامشية، وبوجودها لكل ملصق نستخدم نموذج CRF لاحتساب احتمالية كل حل منها.

التوليد الخطي الموزون (interpolations linear Weighted)

لا يمكن جمع النص والكلام بشكل مباشر على نفس المستوى، فالنص يمكن تطبيعه (Normalize) بأزواج من أحرف العلة الساكنة من أجل معالجتها، فلا يمكن تطبيق ذلك على النسخ الصوتية للنماذج الصوتية، ولذلك ننفذ سجل احتمالات على مستوى الكلمة أنتجه مميِّز الكلام في مقابل استخلاص نتائج معتمدة على مستوى التشكيل من نظام معتمد على النص. وبما أن النتائج الصوتية موجودة ضمن مجال السجل، فإنّ السجلات تُؤخذ من النتائج

المعتمدة على النص، وتكون نتائج النموذجين المعتمدين على الصوت والنص قابلة للإقحام (interpolation).

بعد تنفيذ النظام على مثال ما، اختار المشكّل الآلي المعتمد على النص حلًا بشكل أولي، ولكن بعد الإقحام (interpolation) اختار حلا آخر وهو الحل الصحيح. وأُجريت التجارب باستخدام مجموعة بيانات مقبولة بشكل واسع في النشرات الأدبية، ومن مجموعة البيانات الأساسية المكونة من ٥٢ ألف كلمة اخترنا ١٤٦٥ كلمة منها لتكون بيانات التطوير، أما أوزان التقييم فهي ذاتها المستخدمة في المراجع.

أولاً تستخدم الموارد المعروفة الحالية جميع الرموز (tokens) من مجموعة بيانات لاحتساب الخطأ، ولكننا استثنينا الأرقام وعلامات الترقيم، لأنه ليس لها تشكيل، وثانيًا في حالة احتساب الخطأ دون الحالات الإعرابيّة جرى تضمين الحرف الساكن الأخير من الكلمات في الحساب بشكل تقليدي، ونقترح نحن قياسا جديدا يدعى DERabs no CE وهو قياس يستثني الحرف الأخير، ونعتقد أنه يبيّن بدقة أكبر فرقاً بين أداء أنظمة التشكيل الآلي التي تتضمن الحالات الإعرابيّة وتلك التي تستثنيها.

لقد أمكن الوصول إلى استنتاج أن ميزات النص تؤدي إلى أخطاء في أواخر الكلمات بينما ميزات التسجيلات الصوتية تساهم في الأخطاء على امتداد النص. ووُجد أن معالجة النص نتج عنها أخطاء في آخر الكلمات بنسبة ١٣,٢٪ بينما التسجيلات الصوتية ساهمت بنسبة أخطاء أقل بمقدار ٦,٢٪ وعند استعمالهما سويّة وصلت نسبة الأخطاء إلى ٢,٨٪.

Machine Translation Summit XI, 10-14 September 2007, Copenhagen, Denmark, pp.143-149.

Year of publication: 2007

Arabic Diacritization in the Context of Statistical Machine Translation

Mona Diab, Mahmoud Ghoneim, Nizar Habash

تشكيل اللغة العربيّة في سياق الترجمة الآليّة الإحصائيّة

يدرس هذا البحث تأثير التشكيل في اللغة العربيّة على الترجمة الآليّة الإحصائيّة ويعرّف عددا من طرائق التشكيل التي تتراوح بين التشكيل الكامل إلى الجزئي، ويدرس تأثيرها على الترجمة الآليّة الإحصائيّة ضمن سياقين مختلفين يفرّقان بين تأثير التشكيل على المحاذاة (alignment) وعواقبها على فك التشفير(its consequences on decoding). وفيما يتعلق بتكنولوجيا النطق فإنّ التشكيل الكامل حسّن من جودة أنظمة التمييز الأوتوماتيكي للغة العربيّة (ASR) خاصة المتعلقة باللهجة العاميّة، ولكن ليس هناك دراسة بحثت عن أفضل مستوى تشكيل يكون كافيا لتقديم أفضل النتائج في أنظمة التمييز الأوتوماتيكي للغة العربيّة.

لغرض هذه الدراسة نعرّف عدة تقنيات تشكيل تعبّر عن الظواهر اللغويّة المتعددة الموجودة في النص الطبيعي، فنعالج النص العربي المصدر في سياق الترجمة الآليّة الإحصائيّة المعتمدة على العبارة باستخدام تقنيات التشكيل المتعددة هذه، كما نستكشف اثنين من وسائط المحاذاة، حيث تُستخدم تقنيات التشكيل أو لا تُستخدم لأغراض المحاذاة.

تقنيات التشكيل

عرّفنا ستة مستويات تشكيل واستخدمنا معها جميعا نظامًا لتوضيح اللغة العربيّة. والشكل الموضّح بشكل كامل للكلمة، يُشار إليه بجميع خصائصه الشكليّة ويكون مشكّلا

بشكل كامل. وفي كل تقنية نحذف علامات التشكيل غير المتعلقة بخصائصها أو أهميتها، والتقنيات هي:

دون تشكيل: لا وجود للتشكيل، وتتضمن التشكيل الذي يأتي بشكل تلقائي مع اللفظ (سترمم الجدران).

تشكيل الحرف الأول: التشكيل الإعرابي الذي يكون غالبا مع الأفعال المبنية للمجهول باستخدام الضمة (ستُرمم الجدران).

تشكيل الحرف الأخير: تشكيل إعرابي يُعطى لتوضيح الحالات الإعرابية (سترممُ الجدرانُ).

الشدة: تشكيل معجمي، لوضع الشدة عندما يلزم التضعيف (سترمّم الجدران).

السكون: تشكيل معجمي يُعطي حركة السكون (سترمم الجدْران).

تشكيل كامل: تحديد جميع الحركات (سَتُرَمّمُ الجُدْرَانُ).

استراتيجيات المحاذاة

نتوقع وجود استراتيجيتين تستجيبان بشكل مختلف للتغيير في تقنيات التشكيل، الأولى هي المحاذاة الأساسية، فنطبّق محاذاة للكلمات باستخدام بيانات تدريب مُستخدمة من إحدى تقنيات التشكيل، أما الثانية فهي المحاذاة مع إعادة الترتيب، تُطبّق المحاذاة مع تقنية التشكيل الأولى ثم نصل بين النص المصدر في المحاذاة والنص ذي التشكيل الذي يهمّنا.

الاستراتيجية الثانية عند تطبيقها مع أيّ من تقنيات التشكيل الأخرى ستعاني من التشتت (sparsity) بسبب تزايد عدد أنواع الكلمات، والتقنية الأولى من التشكيل تكون أقوى فيما يتعلق بالمحاذاة لأن عدد أنواع الكلمات أقل مقارنة بغيرها من التقنيات. وعند الربط بين المحاذاة دون تشكيل وتقنية التشكيل المرغوبة في الاستراتيجية الثانية تتحقق فائدة حصولنا على عبارات في الجدول أكثر من تلك التي نحصل عليها في الاستراتيجية الأولى لنفس تقنية التشكيل، فمثلا بمقارنة جدول العبارات لتقنية التشكيل الأولى (دون تشكيل) مع المحاذاة الأساسية الكاملة،

فإنّ حجم جدول العبارات يزيد بنسبة ٦٦ ,٦٪ ويقل عدم الوضوح، ويقلّ عدد العبارات المصدرية المرتبط بعبارات الهدف من ٤٢٦ ,١ لعدم التشكيل إلى ٤٠٧ ,١ للمحاذاة الأساسية الكاملة . بينما في المحاذاة مع إعادة الترتيب الكاملة تكون هناك زيادة بنسبة ٦٣ ,٣٪ في حجم جدول العبارات استجابة لعدم وضوح بنسبة ١٤٠ ,١ .

فإن استخلاص جدول العبارات وتشفيرها في الاستراتيجيتين يستمر بأسلوب الترجمة الآليّة الإحصائيّة النمطية المعتمدة على العبارة. وبيانات الاختبار والضبط تُوضّح بنفس تقنية التشكيل المستخدمة في بيانات التدريب في كلا الاستراتيجيتين.

البيانات التجريبية

نستخدم الأخبار باللغتين العربيّة والإنجليزية المكوّنة من خمسة ملايين كلمة كبيانات تدريب لنموذج الترجمة، وتتم مرحلة ما قبل المعالجة للغة الإنجليزية بتحويل الكلمات إلى الأحرف الصغيرة، وفصل علامات الترقيم عن الكلمات، وفصل علامة الجمع "'s" وتتم العملية نفسها لجميع التجارب، أما ما قبل المعالجة للعربية فتختلف فيما يتعلق بالتشكيل فقط ، ونفترض نفس عملية التقطيع (tokenization) لجميع البيانات العربيّة.

نظام الترجمة الآلية الإحصائية

استخدم في جميع التجارب نظام ترجمة آليّة إحصائية معتمد على العبارة (phrase-based) من الأنظمة المتوفّرة تجاريا، واعتمدت الأوزان بمقياس تقييم غير حساس لنوعية الكلام. وقمنا في كل تقنية بتدريب نظامين لكل استراتيجية محاذاة.

النتائج

أسوأ النتائج تكون في حالة استخدام تقنية التشكيل الكامل لكلا استراتيجيتي المحاذاة، ويليها تقنية «تشكيل الحرف الأخير»، والفرق بينهما هو نقطة واحدة على الأقل للاستراتيجيتين

ولجميع مجموعات الاختبار، وجميع شروط « تشكيل الحرف الأول « تتفوق على تقنية «دون تشكيل» لجميع مجموعات الاختبار عدا واحدة. والفرق بينهما غير مهم إحصائيا، إلا أن تقنية «السكون» تتفوق على تقنية «عدم التشكيل» لجميع مجموعات الاختبار مع أن الفرق غير مهم إحصائيا.

بعد اختبار فائدة تطبيق تشكيل كامل أو جزئي في سياق نظام ترجمة آلية إحصائية معتمد على العبارة، فإن النتائج تؤكد أن التشكيل الكامل غير مفيد في الترجمة الآلية الإحصائية، وبجميع الأحوال لا تختلف تقنيات التشكيل الأخرى بشكل كبير عن الحالة الأساسية من عدم التشكيل، وفي الواقع فإن اثنتين من تقنيات التشكيل الجزئي « تشكيل الحرف الأول» و»سكون» كان أداؤهما أفضل قليلا من «عدم التشكيل» في اثنتين من مجموعات الاختبار. بالتالي يقترح بحثنا أن أداء الترجمة الآلية الإحصائية يرتبط إيجابا بعدد المقاطع (tokens) المتأثرة بشكل دقيق بتقنية التشكيل إلى درجة أنها قد تجعلها قوية لنسب الكلمات غير المعروفة (Out of Vocabulary) المرتفعة قليلا. إن هناك حاجة لأبحاث تستهدف ظواهر تشكيل محددة مثل المبني للمجهول (passivization) وتشكيل التوضيح المعجمي مثل الشدة. إن مثل هذه الأبحاث مطلوبة أولا قبل أن نرى تأثيرا كبيرا على تطبيقات معالجة اللغات الطبيعية مثل الترجمة الآلية الإحصائية.

International Journal on Document Analysis and Recognition (IJDAR), June 2015, Volume 18, Issue 2, pp 183–197 (IJDAR (2015) 18: 183)

Year of publication: 2015

Automatic diacritization of Arabic text using recurrent neural networks

Gheith A. Abandah, Alex Graves, Balkees Al-Shagoor, Alaa Arabiyat, Fuad Jamour, Majid Al-Taee

التشكيل الآلي للغة العربيّة باستخدام الشبكات العصبيّة المتكررة

يقدّم هذا البحث طريقة للتشكيل الآلي للنصوص العربيّة، باستعمال الشبكات العصبيّة المتكررة (recurrent neural network (RNN))، وهي مصممة لتحويل النص العربي غير المشكّل إلى جمل مشكّلة بشكل كامل، ونستخدم شبكة بذاكرة قصيرة الأجل عميقة وثنائية الاتجاه (LSTM) تبني اختصارات للنصوص عالية المستوى تستغلّ السياق واسع المدى في كلا جهتي الإدخال، وتختلف هذه الطريقة عن سابقاتها بأنها لا تقوم بتحليل معجمي أو شكلي أو نحوي على البيانات قبل معالجتها، ومع ذلك عند معالجة الشبكة لاحقا بتقنياتنا لتصحيح الأخطاء فإنها تحقق تشكيلا معقولا، ونسبة خطأ تتراوح بين ٠٩ ,٢٪ و ٨٢ ,٥٪ على عينات من ١١ كتابا، وبذلك تقلل هذه الطريقة الخطأ في التشكيل بنسبة ٢٥٪، وفي الكلمات بنسبة ٢٠٪، وفي الحرف الأخير بنسبة ٣٣٪ على أفضل النتائج المنشورة.

إن هذه الطريقة الإحصائية الخالصة من النسخ المتسلسل (sequence transcription) تعتمد على بناء long term-short memory (LSTM)) ثنائي الاتجاه العميق (deep bidirectional) ، وقد طُبِّق بنجاح على العديد من مهمات النسخ المتسلسل، منها تمييز الكلام وتمييز خط اليد الأوتوماتيكي. وحسب علمنا فإن هذا العمل هو أول من يستخدم RNN للنسخ المتسلسل لإضافة التشكيل للنصوص العربية بشكل أوتوماتيكي. وقد جُرِّب باستخدام البرمجية مفتوحة المصدر RNNLIB.

النسخ المتسلسل

إن الشبكة العصبية المتكررة الأساسية المستخدمة هنا هي **LSTM** ثنائية الاتجاه العميقة، وجمع هذا البناء ذاكرة قصيرة الأمد طويلة (**LSTM**) مع شبكات عصبية متكررة ثنائية الاتجاه، والطبقات المخفية المتراصة الموجودة في الشبكات العصبية ذات التغذية الأمامية. طُبِّقت شبكات **LSTM** سابقًا على النص التنبؤي على مستوى الحرف، وأثبتت قدرتها في تمثيل ملحقات لغوية طويلة المدى.

ذاكرة قصيرة الأمد طويلة

إن هذه الذاكرة التي تستخدم خلايا ذاكرة مبنية لهدف هي أفضل في إيجاد واستخدام السياق طويل المدى.

الشبكات العصبية المتكررة ثنائية الاتجاه (BRNNs)

هناك عيب واحد في الشبكات العصبية المتكررة التتقليدية، وهو أنها قادرة على الاستفادة من السياق السابق فقط. أما في التشكل الآلي حيث تُنسخ الجملة كاملة دفعة واحدة فليس هناك سبب لعدم الاستفادة من السياق المستقبلي أيضًا، والشبكات العصبية المتكررة ثنائية الاتجاه تفعل ذلك من خلال معالجة البيانات من الاتجاهين بطبقتين مخفيتين منفصلتين وتقدمان تغذية لنفس طبقة المخرجات.

والجمع بين **BRNNs** و **LSTM** يعطينا **LSTM** ثنائي الاتجاه، يمكن أن يصل إلى سياق طويل المدى في كلا اتجاهي الإدخال.

الشبكات العصبية المتكرر العميقة (Deep RNNs)

العنصر الأهم في النجاح الذي حققته الأنظمة المهجنة مؤخرا هي الهيكليات العميقة التي يمكنها أن تبني تمثيلات بمستوى أعلى للنص، وتُبنى هذه الشبكات بوضع عدة طبقات من الشبكات العصبية المتكررة فوق بعضها، فتمثل مخرجات كل طبقة مدخلات للطبقة التالية، على

افتراض أن نفس اقتران الطبقة الخفية مستخدم في جميع الطبقات. وهناك طريقتان تُستخدمان لتدريب الشبكات العصبية المتكررة لنسخ تسلسل من حروف اللغة العربية غيرالمشكّلَة، مع نظيراتها من الأحرف المشكّلَة، وهي:

شبكة واحد إلى واحد «one-to-one»

نستخدم هنا تشفير الحروف واحد إلى واحد، للحرص على أن تكون متواليات الهدف لها ما يقابلها بشكل واحد إلى واحد من متواليات المدخلات، وقد دُرّبت الشبكة لتصنيف كل حرف مدخل على حدة مع ما يقابله من النسخة المشكّلة.

شبكة واحد إلى أكثر "many-to-one"

تنسخ هذه الشبكة المدخلات غير المشكّلة بعد تشفيرها باستعمال **Unicode** للمخرجات المشكّلة باستخدام تشفير «واحد إلى أكثر"، وهنا تكون متواليات المخرجات أطول من متواليات المدخلات.

الطريقة الثانية تستخدم **LSTM** إضافية من طبقة واحدة للتنبؤ برموز مخرجات «الواحد إلى أكثر " وتستند هذه الطريقة على محول الطاقة المتسلسل (**sequence transducer**) القادر على إنتاج مخرجات على شكل واحد وصفر أو عدة مخرجات على شكل ملصقات لكل ملصق مدخل، كما أن مولد الطاقة قادر على ترتيب متواليات المدخلات (**input sequence**) والمخرجات، أي لا تكون هناك حاجة لتحديد أي تشكيل يقابل كل حرف بشكل مسبق. ويستخدم محول الطاقة اثنتين من **RNN** منفصلتين: شبكة المدخلات ثنائية الاتجاه، وشبكة التنبؤ التي لا بد أن تكون أحادية الاتجاه، وشبكة مخرجات تغذية أمامية تُستخدم لجمع مخرجات الشبكتين.

بناء تجريبي

لأغراض التقييم جرى تشفير الجملة العربية المشكّلة في سجل ملائم للنسخ المتسلسل للشبكات العصبية المتكررة، وهذا السجل يتكوّن من التسلسل الهدف المشكّل وتسلسل المدخلات غير المشكّلة، فتنسخ الشبكات العصبية المتكررة المدخلات إلى مخرجات مشكّلة بشكل كامل. من ثم نطبّق تصحيحات بعد المعالجة لتصحيح أي أخطاء قد تظهر، وفي النهاية نقارن المخرجات المصححة بالهدف لمعرفة دقة التشكيل.

خلال تدريب **RNN** تعرض المدخلات والتسلسل الهدف على الشبكة، لتشكيل جملة غير مشكّلة يكون الحقل الهدف غير متوفر وبالتالي لا تتم المقارنة.

أُخذت البيانات التجريبية المستخدمة من الكتب العشر لمجموعة التشكيل من كتب التراث الديني الإسلامي، مع جزء بسيط من القرآن الكريم، وكُتبت هذه الكتب بعربية فصيحة بتشكيل كامل بصيغة **HTML**، ولكن الكتاب الأول مشكّل جزئيا واختير لدراسة أثر استخدام المدخلات المشكّلة جزئيا على دقة التشكيل. كما استخدمنا **LDC›s Arabic Treebank** من قصص أخبار مشكّلة، أحدها (**LDC ATB٣**) يعد مثالا للغة العربية الحديثة ويتألف من ٥٩٩ خبرا من مطبوعة النهار اللبنانية، واستخدمت في هذا العمل في التجارب النهائية لتسهيل المقارنة مع الأنظمة القديمة.

بعض هذه الكتب الكلاسيكية كبيرة وتأخذ وقتا طويلا للمعالجة، واخترنا بشكل عشوائي مجموعات فرعية من جمل كل كتاب للحصول على مجموعات بيانات مختلفة في الحجم، فاخترنا مثلا ٣٪ من الكتاب الأول في النهاية و١٠٠٪ من الكتاب الرابع و **ATB٣** في النهاية الثانية. نحضّر الكتب للتدريب والاختبار بتحويلها من **HTML** إلى ملفات نصوص عادية فيها جملة في كل سطر، والقرآن الكريم تكون كل آية في سطر.

نستخلص النسخة المشكّلة من جمل **ATB٣** من تنسيقها المتكامل، حين تكون كل جملة في سطر، والكلمات المشكّلة في هذا الشكل متوفرة في الترجمة الصوتية العربية **Buckwalter**، ولأن بعض الكلمات غير متوفرة في النسخة المشكّلة نستخدم المصدر غير المشكّل لهذه الكلمات. من ثم تُحول جميع أسطر البيانات إلى تشفير يونيكود.

تشفير البيانات

تُحوّل الجمل إلى سجلات جمل كل منها يحتوي نسختين: تسلسل مدخلات دون تشكيل، وتسلسل هدف مع التشكيل مفصولة بفاصلة، والنسخة غير المشكلة تُستخلص من المشكّلة بعد إزالة التشكيل. تُمثل علامات التشكيل في يونيكود كحروف إضافية، مثلا: كلمة «ثم» لها سجل من حقلين «ثم» ، «ثُمَّ» وتُشفر على شكل « ٠٦٢B ٠٦٤٥" ، « ٠٦٢B ٠٦٤F ٠٦٤٥ ٠٦٥١ ٠٦٤E" وبالتالي تسلسل الهدف أطول من تسلسل المدخلات، وهذا التشفير «واحد إلى أكثر" يُستخدم مع شبكات "واحد إلى أكثر".

أما في شبكات «واحد لواحد» فنشفر كل تشكيل محتمل لكل حرف برمز واحد، وبالتالي تكون المدخلات والمخرجات بنفس الطول.

معاملات التدريب

تُدرب الشبكات العصبية المتكررة في هذا العمل ب ٨٨٪ من الجمل المتوفرة، وما تبقى يُختار بشكل عشوائي لأغراض الاختبار. سنستخدم نفس المجموعة للتطوير والاختبار، مع العلم أن استخدام مجموعة الاختبار كمجموعة تحقق في تدريب الشبكات العصبية سيعطي شبكة ملائمة لمجموعة الاختبار، كما أننا جربنا استخدام مجموعات تطوير واختبار مختلفة، وهنا جمل التدريب تقسم بشكل عشوائي إلى نسبى ٧٠٪ للاختبار و٣٠٪ للتحقق، وتُستخدم الشبكة الناتجة لإيجاد دقة التشكيل.

المعالجة اللاحقة

نستخدم تقنيات معينة لتحسين المخرجات، مثل تصحيح الأحرف إذا ما حصل تغير في أحد أحرف الكلمة، وحذف السكون، ووضع الفتحة على الأحرف التي تسبق الألف المقصورة أو التاء المربوطة في حال ظهور أي حركة أخرى، ومقارنة الكلمة المشكّلة النهائية مع القاموس الذي يحوي جميع احتمالات التشكيل للكلمة، فإذا لم تكن الكلمة موجودة فيه تُقرَّب إلى أقرب احتمال ممكن منها.

The Arabian Journal for Science and Engineering, Volume 35, Number 2C

Year of publication: 2010

Automatic restoration of Arabic diacritics: A simple, purely statistical approach

Mansour Alghamdi, Zeeshan Muzaffar, and Hazim Alhakami

استعادة أوتوماتيكية لعلامات التشكيل العربيّة: منهجية بسيطة وإحصائية بحتة

تعرض هذه الورقة طريقة مبتكرة لتشكيل النص العربي آليا، حيث تعتمد معظم الطرق السابقة المنشورة على أدوات أخرى مثل أدوات أنموذج ماركوف الخفي و/ أو المعرفة اللغوية مثل الصرف والنحو، والنظام الذي بين أيدينا هو نظام أساسي يتميز بصغر حجمه وسرعته في المعالجة بمعزل عن الأدوات الحاسوبية المعروفة أو القوانين اللغوية، إذ يستخدم هذا النظام طريقة إحصائية بحتة تعتمد على احتمالية التسلسل الرباعي للحروف، وينتج عنه درجة عالية في دقة التشكيل إذا ما قورن بالأنظمة السابقة التي تعتمد على الإحصاء فقط بوصفه نظاما أساسيا.

المنهجيات والإجراءات

إن التقنية المستخدمة في هذا النظام فيها خطوتان أساسيتان، الخطوة الأولى هي صياغة قائمة غنية جدا من الأنماط الرباعية، وهي (الأنماط المكونة من أربعة أحرف مشكّلة متتالية) المستخدمة باستمرار، والخطوة الثانية هي باستخدام هذه القائمة لتشكيل أي نص عربي تقريبا.

استخدمنا النص العربي المشكّل الذي يُدعى» **KDATD"** (وهو نص جمعه فريق من مدينة الملك عبدالعزيز للعلوم والتقنية) لصنع هذه القائمة، ويتكون هذا النص من ٢٣١ ملف نصيٍّ يمثل ٢٢ موضوعا، كل ملف منها يحتوي على ما يقارب ١٠٠٠ كلمة مشكّلة، فقد استُخلصت الأنماط الرباعية من هذا النص مع تكراراتها، كما أن المسافة بين الكلمات جرى أخذها بعين الاعتبار في الأنماط الرباعية، ووضعت الأنماط التي تحتوي على نفس تسلسل الأحرف، ولكن

بعلامات تشكيل مختلفة في مجموعات، واحتسبت احتمالية ظهور كل عنصر من كل مجموعة، ثم اختيرت الأنماط ذات الاحتمالية الأعلى من كل مجموعة، واحتسبت احتمالية النمط الرباعي الذي يملك أعلى تكرار للأحرف وعلامات التشكيل معا، وكانت النتيجة عبارة عن قائمة (قاعدة بيانات) مكونة من ٦٨٣٧٨ نمطٍ رباعيٍّ يملك أعلى احتمالات.

طُوِّر النظام لتشكيل النص العربي غير المشكّل باستخدام قاعدة بيانات الأنماط الرباعية؛ إذ تكون مدخلات النظام سلسة من الكلمات، أي جمل، ويفترض النظام أن كل جملة هي سلسلة من الأحرف غير المشكّلة، والهدف هو تشكيلها بتطبيق قاعدة البيانات المحتسبة سابقا.

عندما يكون لدينا سلسلة من الأحرف غير المشكّلة فإننا نضيف مسافة قبلها وبعدها، ثم نقسمها على شكل أنماط رباعية كما يلي:

لكل نمط منها نبحث عن النمط المشكّل المقابل له في قاعدة البيانات، وفي هذه المرحلة ومن أجل الحفاظ على السهولة فإننا نفترض أن جميع المتتاليات موجودة، فنستخلصها ويصبح لدينا علامة تشكيل لكل حرف في كل نمط بالإضافة إلى احتمالية ورود كل نمط منها.

تكون الخطوة التالية بتشكيل كل حرف عن طريق الأخذ بعين الاعتبار جميع سلاسل الأنماط الرباعية المتتالية التي تحتوي ذلك الحرف من قاعدة البيانات، وبالطبع هناك أربعة أنماط رباعية مشكّلة على الأكثر لكل حرف. ولكل حرف سنعتبر جميع حالات ظهوره في الأنماط الرباعية المتتالية عن طريق جمع احتمالات الظهور لنفس علامة التشكيل لكل الحروف، في النهاية تُختار علامة التشكيل المرتبطة بالنمط الرباعي الذي له أعلى مجموع احتمالات.

إذا كان هناك سلسة نمط رباعي مشكّلة غير متوفرة في قاعدة البيانات فإن النظام يضع فراغا في مكانها، مثلا: إذا كان لدينا سلسلة الأحرف ق ب ج د ن ل م ف ص ط، فإن الأنماط المتوقعة ستكون م ف ص ط – ل م ف ص – ن ل م ف – د ن ل م– ج د ن ل – ب ج د ن – ق ب ج د، ولنفترض أن النمط الثاني لم يكن موجودا في قاعدة البيانات فسيضع النظام مكانه فراغا ثم يكمل باقي الأنماط.

كانت نتيجة هذه العملية عبارة عن نظام حجمه ثلاث وحدات ميجابايت وبلغت سرعة

المعالجة أكثر من ٥٠٠ كلمة في الثانية باستخدام معالج ٥٣٣ ميغاهيرتز.

يُظهر اختبار النظام أن هناك نسبة خطأ تصل إلى ٤٦ ,٧٪ عندما كانت مجموعة الاختبار (مكونة من ٥٠١٧ كلمة) من مجموعة التدريب، وكانت نسبة الخطأ ٥٢ ,٨٪ عندما كانت مجموعة الاختبار (مكونة من ٣٨٠٠ كلمة) غير مضمنة في مجموعة التدريب.

لاختبار النظام على قاعدة بيانات معروفة، اختير ٦٤ ملفا من قاعدة بيانات معروفة، تضم هذه الملفات ١٥٨٣ كلمة و٨٣٨٩٧ حرفا، ولكن احتوت هذه القاعدة على بعض الأخطاء والكلمات الأجنبية وأسماء الأعلام، ونتيجة لذلك ارتفعت نسبة الخطأ، لأن مجموعة الاختبار ليست من نفس مصدر مجموعة التدريب مثل الدراسات السابقة. وما يجعل هذه الطريقة مختلفة عن سابقاتها أنها لا تتوقف عند نمط رباعي واحد لتشكيل حرف ما، ولكنها تجمع أربعة أنماط رباعية مختلفة، كما أن وجود الأحرف المتتالية يزيد من احتمالية الدقة، لأنه مع زيادة الأحرف المتتالية تزيد دقة التشكيل.

كما هو معلوم، إذا كانت علامات التشكيل لجميع أحرف الجملة معروفة فستكون لدينا عملية تشكيل خالية من الأخطاء، أما إذا اعتمدنا على احتمالية النمط الأحادي فإن معدلات الدقة ستكون منخفضة للغاية، بالإضافة إلى أن النظام لا يعمل على قائمة محددة مسبقا، فيمكنه تشكيل أي كلمات عربية سواء أكانت في مجموعة التدريب أم لا.

إن التقييم الأولي للنظام مشجع، ولكن يمكن تحسين نسبة الدقة أكثر بإضافة الأنماط الرباعية الأخرى المحتملة غير المتوفرة في KDATD، بالإضافة إلى أنه يمكن تغذية النظام بمعلومات لغوية بإضافة قواعد نحوية وشكلية من شأنها أن تحسن معدلات الدقة.

The Proceedings of the 12th European Chapter of the Association for Computational Linguistics (EACL 2009) Workshop on Computational Approaches to Semitic Languages, 30 March - 3 April 2009, Athens, Greece.

Year of publication: 2009

A Hybrid Approach for Building Arabic Diacritizer

Khaled Shaalan, Hitham M. Abo Bakr, Ibrahim Ziedan

طريقة هجينة لبناء مشكّل عربي

في هذا البحث قمنا بإضافة التشكيل إلى نص غير مشكول باستخدام طريقة هجينة، فهذه الطريقة تحتاج إلى معجم لغوي عربي وذخيرة كبيرة لنصوص مشكولة كاملة لتحديد الحركات، وقد اعتُبِرت معالجة تشكيل نهايات الكلمات عملا مستقلا باستخدام معلومات نحْويّة. كما اعتمد البحث على الاسترجاع المعجمي والمركّبات المزدوجة (bigram) وتقنيات SVM الإحصائية الأولية.

لقد حُلَّت مشكلة الاسترجاع الآلي لإشارات التشكيل للنص العربي بطريقتين:

الأولى منهج يعتمد على القواعد، ويتضمن تكاملاً معقّداً بين الأدوات الصرفيّة واللغويّة والنحْويّة بجهود كبيرة لتحصيل قواعد لغويّة متخصصة. والمحلل الصرفي هنا يقوم بترتيب فواصل (breakdowns) الكلمات غير المشكولة حسب أنماط وقوالب معرّفة (patterns and Templates) ويتعرف على السوابق واللواحق (prefixes and suffixes). أما محلل البنية اللغويّة (syntax analyzer) فيطبق قواعد لغويّة متخصصة لتحديد تشكيل آخر الكلمات (case-ending) عادة تستخدم فيها قواعد الحالة المنتهية (finite state automata).

المنهج الثاني هو منهج إحصائي يتطلّب ذخائر وموارد لغويّة موسومة (large tagged corpus) لاستخراج إحصائيات لتوقّع علامات التشكيل المفقودة ولا يحتاج إلى معرفة لغويّة مسبقة.

هذه الدراسة قدمت نموذجاً بسيطاً لكنّه فعّال لتحصيله نتائج مقاربة لأفضل الأنظمة الشبيهة، فالنتائج المنجزة هي: ١١,٧٩٥٪ نسبة خطأ الكلمة (WER)، و٣,٢٤٥٪ نسبة خطأ التشكيل (DER).

نموذج التشكيل المقترح:

هناك ثلاث تقنيات افتُرضت لحل مشكلة التشكيل في العربيّة، لكلّ منها نقاط قوة ونقاط ضعف، وهي كما يأتي:

الاسترجاع من المعجم (lexicon retrieval): هذه الطريقة تحاول إيجاد الكلمات المشكولة من ذخيرة معجميّة مكوّنة من كلمات غير مشكولة، فإذا رجعت كلمة مشكولة واحدة فهو الحل النهائي، ولا نحتاج إلى الرجوع إلى النتائج مرة أخرى.

مقاطع ثنائية مشكولة (diacritized bigram): هذه الطريقة تأتي عندما يُسترجع أكثر من نتيجة من الطريقة السابقة، ويُستخدم تعبير مكون من كلمتين (bigram) حيث تُقرر إحدى الكلمتين الحالة الإعرابية للكلمة الأخرى.

مشكّل مبني على الإحصاء (SVM statistical based diacritizer): الطريقتان السابقتان صُنّفتا بعمليتي بحث إما للبحث عن كلمة في المعجم أو عن تعبير مزدوج الكلمات في قاعدة بيانات الكلمات المزدوجة (bigram database). الفكرة الرئيسة لهذه الطريقة هي عنونة الرموز بأوسمة أقسام الكلام (POS tags) ثم يبحث في المعجم باستخدام الرمز (token) والقسم المقابل من أقسام الكلام (POS)، فيُستخرج التشكيل الصحيح والكلمات المزدوجة الغامضة من المعجم.

تجريب المشكّل العربي:

جُرِّب وفُحص نظام التشكيل على مقالات إخبارية مشكولة، وشملت الذخيرة اللفظ الكامل شاملاً نهايات الكلمات، فنتج من ذلك في النهاية جزء واضح وصحيح لقسم من

الذخيرة، لذا يمكن أن تُدرس النتائج وتُطور بالمستقبل لزيادة الدقة والصحة.

الخلاصة والعمل المستقبلي:

في هذه الدراسة افترضنا نظام تشكيل يميّز بين الحركات ضمن الكلمة ونهاية الكلمات، فكانت النتيجة الإجمالية مقاربة لأفضل النتائج السابقة في هذا المجال، وقد أظهرت الطرق الإحصائية نتائج جيدة لحل غموض التشكيل، وقدّم النظام المفترض نتائج جيدة بين DER, WER ومقارنتهما مع نظام MADA-D، وقد حسّنت التعديلات التي جرت لخوارزمية نهاية الحالة أداء النتائج.

ونتطلع مستقبلا إلى إضافة مجموعة من القواعد الإرشادية والتحضيرية، وبالتالي تحسين الأداء لاسترجاع التشكيل الصحيح لضمائر الملكية، وإضافة رموز POS إضافية، للتمييز بين المثنى والجمع.

The 2nd International Conference on Arabic Language Resources and Tools

Year of publication: 2009

Hybrid System for Automatic Arabic Diacritization

Mohsen A. A. Rashwan, Mohammad Al-Badrashiny, Mohamed Attia, Sherif M. Abdou

نظام هجين للتشكيل العربي الآلي

تقدّم هذه الورقة نظاما من طبقتين لتشكيل اللغة العربيّة، تحاول الطبقة الأولى أن تحدد علامات التشكيل الأكثر احتمالية، أما الطيقة الثانية فتحلل كلّ كلمة عربيّة إلى مكوناتها الشكلية. وتشكيل الكلمة العربية يتكوّن من مكونين، أحدهما يعتمد على الشكل والثاني على المحل النحوي.

التحليل الشكلي

نموذجنا الشكلي يقسّم الكلمات العربيّة إلى أربعة أجزاء، هي سابقة وجذر ووزن ولاحقة (مثلا: تتناوله: التاء هي السابقة، نول الجذر، تفاعل هو الوزن، والهاء هي اللاحقة)، وكلّ كلمة تنتمي إلى واحدة من أربعة أصناف هي مشتقات عادية، مشتقات غير عادية، ثابت، معرب.

تحديد القسم من الكلام

يعتمد نموذج تحديد القسم من الكلام (POS Tagging) مجموعة من أقسام الكلام تتكون من ٦٢ علامة فقط، تُعطى لكل كلمة العلامات المناسبة (مثال: الكتابات، ال التعريف، اسم، جمع، مؤنث).

التشكيل من خلال فك اللبس (Disambiguating) الإحصائي للنص العربي المحلل يمكن أن يُستخلص التشكيل من السابقة والوزن واللاحقة، أي من التحليل الشكلي

للكلمة، ولكن تكمن المشكلة في توضيح التحليلات المتعددة التي يقترحها المحلل الشكلي، لذلك يجب اختيار التحليل الأكثر احتمالا. أما التشكيل النحوي فإنه يستخلص أقسام الكلام لسلسلة من الكلمات العربية مع التشكيل النحوي لكل كلمة بعد فك اللبس صرفياً (morphological disambiguation)، ثم يُطبّق التوضيح النحوي مرة أخرى لمعرفة التسلسل ذي التشكيل النحوي وأقسام الكلام الأكثر احتمالا.

المزج بين الكلمات الكاملة والمحللة

لتحسين أداء نظام تشكيل النص المحلل، طوّرنا نظاما هجينا يجمع بين نظام التشكيل المعتمد على الشكل، ونظام آخر يعتمد على الكلمات كاملة. ونستخدم قاموسا كاملا من الكلمات العربيّة مع شروحاتها الصوتية، ويبحث النظام عن كلّ كلمة مدخلة في القاموس، فإذا وجدت مثل هذه الكلمة سميت «كلمة قابلة للتحليل» (analyzable) وتُستخرج جميع احتمالات التشكيل لها وهذا يسمى تحليل الكلمة، أما سلسلة الكلمات القابلة للتحليل فتسمى «المقطع القابل للتحليل»، ويكون تحليل كلمات هذا المقطع شبكة تُوضّح لمعرفة تسلسل التشكيل الأكثر احتمالا، ثم توصل الكلمات الكاملة المشكّلة من المقاطع القابلة للتحليل مع الكلمات المدخلة غير القابلة للتحليل - إن وجدت- للحصول على نص عربي أكثر وضوحا، وهذا التسلسل الأخير يُدخل إلى النظام المذكور سابقا (نظام التوضيح والتحليل).

طريقة التوضيح الإحصائي

إن اللبس (ambiguity) في الحلول العديدة الممكنة لكل كلمة تقود إلى وجود شبكة من التحليلات الممكنة، ولحل هذا الإبهام فإننا نعتمد على طريقة تقدير الاحتمالية الاستدلالية القصوى (most statistically sound sequence).

البناء التجريبي

استخدمت قاعدة بيانات لتدريب النظام سابق الذكر، وتتكوّن قاعدة البيانات من نص

عربي بحجم يقارب ٧٥٠ ألف كلمة محللة شكليا وممنوحة علامات أقسام الكلام ومشفرة صوتيا، وذخيرة لغوية إضافية مكونة من ٢٥٠٠ ألف كلمة مشفرة صوتيا (**phonetically transcripted**) فقط، وبيانات اختبار مكونة من ١١ ألف كلمة مشروحة يدويا (**manually annotated**) من ناحية التشكيل ومن حيث أقسام الكلام والمقاطع الصوتية.

بعد إجراء عدة تجارب للنظام وجد أن أفضل استراتيجية هي الجمع بين المنهجيات الإحصائية ومنهجيات الاستدلال اللغوي مثل التحليل الشكلي وعلامات أقسام الكلام.

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 65–72, October 25, 2014, Doha, Qatar.

Year of publication: 2014

Automatic Arabic diacritics restoration based on deep nets

Mohsen A. A. Rashwan, Ahmad A. Al Sallab, Hazem M. Raafat, Ahmed Rafea

الاسترجاع التلقائي للتشكيل في اللغة العربيّة اعتمادا على الشبكات العميقة

هناك نوعان من التشكيل؛ الصرفي الذي يُعنى بمعنى الكلمة، والنحوي الذي يعنى بموقعها من الإعراب، يؤدي تطبيق التشكيل الصرفي إلى نسبة خطأ قليلة جدا، لذلك نركز هنا على التشكيل النحوي مستخدمين الشبكات العصبية العميقة (**deep neural nets (DNN)**)، والخصائص التي ستُختبر هي القسم من الكلام، والرباعيات الصرفية (**morphological quadruple**)، والحرف الأخير، وهوية الكلمة.

يجري إدخال المدخلات إلى النظام كلمة كلمة، وتُعامل مع السياق بتخزين ثلاث كلمات سابقة وكلمة لاحقة لكل كلمة، وإذا كانت الكلمة هي الأخيرة أو الأولى في الجملة فإنه يُستعاض عن السياق السابق أو اللاحق بصفر، ويتركز الاهتمام على سياق الكلمات في التشكيل النحوي وعلى سياق الأحرف في التشكيل الصرفي.

يعتمد استخلاص الخصائص على الخصائص نفسها، فلأقسام الكلام ندرّب شبكة عصبية عميقة خاصة، أما الخصائص الأخرى فتُستخلص مباشرة من الكلمة، وحالما تكون الخصائص جاهزة لتسلسل معين من الكلمات فإنها تدخل إلى مصنِّف شبكة عصبية عميقة، للحصول على تقدير أولي للتشكيل المطلوب، ويجري إيجاد التشكيل النهائي في مرحلة «فك اللبس للفئات الفرعية » (**Confused Sub-Classes Resolution CSR**).

يعتمد فك اللبس للفئات الفرعية **CSR** على تحليل مصفوفة اللبس (**confusion matrix**)

وتكون مخرجات هذا التحليل عبارة عن شبكة مكونة من مصنِّفات فرعية هدفها فك اللبس الذي يظهر من تحليل مصفوفة اللبس. وتبدأ هذه الطريقة بتدريب مصنف شامل ثم تقيِّم أداءه، ولتحسين الدقة يُحَلل مصدر الأخطاء عن طريق بناء مصفوفة اللبس من أجل مجموعة التدريب، وموقع العنصر القطري الذي يحدد زوج الفئات الملتبسة مع بعضه.

تبدأ آلية عمل هذه المنهجية بتدريب مصنف شامل أساسي في الشبكة العصبية العميقة للحصول على مصفوفة اللبس على مجموعة التدريب، ثم تحديد مجال اللبس الذي يكون أعلى من نقطة عتبة محددة نحصل عليها من تحليل مصفوفة اللبس، ثم تدريب مصنفات فرعية لكل مجال فيه لبس، ثم تحديد بناء النموذج الذي له مصنفات فرعية عند إشارة معينة في الطبقة وعند عمق الطبقة التي يُحل عندها هذا المجال. وتكون مدخلات مرحلة معالجة النصوص عبارة عن ملف يحتوي النص الخام، وقد ثبت أنه يمكن الاستفادة من مجموعة من الخصائص في تشكيل النصوص العربية، وهذه الخصائص هي: هوية الحرف الأخير، وهوية الكلمة الخام، وخصائص السياق، وعلامات أقسام الكلام.

بعد دراسة تأثير منهجية فك اللبس للفئات الفرعية، بينت النتائج وجود تحسّن، كما وُجد تحسن في النتائج عند دراسة تأثير التعلم من السياق وتأثير خاصية تشكيل الحرف الأخير، ونجد عند مقارنة هذا النظام بغيره من الأنظمة السابقة وجود تحسن في النتائج.

دُرِست في هذا البحث عمليّة تشكيل اللغة العربيّة تحت تأثير إطار التعلم العميق والاستفادة من تدريب نماذج الشبكات العصبية العميقة، كما استخدم محدد لعلامات أقسام الكلام (POS tagger)، وجرى تحسين دقة النتائج عن طريق استخدام فك اللبس للفئات الفرعية ، والاستفادة من خصائص النص المذكورة سابقا.

٣-٤-٥ أبحاث البحث في النصوص

وتضم ٦ أبحاث من نوع (ب) هي: التحديات في استرجاع المعلومات من البيانات العربيّة غير المهيكلة، وسمة تقييم البحث الحاسوبي في القرآن، والتنقيب في بيانات الحديث

النبوي وتصنيفه: دراسة مقارنة، وتوسيع الاستعلام القائم على الأنطولوجيا لاسترجاع النصوص العربيّة، و محرك بحث مقترح للوثائق العربيّة القديمة (المخطوطات)، وأسلوب تخصيص عنقدة ماركوف والتعليم العميق لتصنيف النصوص العربيّة.

2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, UK, pp. 456-461.

Year of publication: 2014

Challenges in Information Retrieval from Unstructured Arabic Data

Hussein Khalil, Taha Osman

التحديات في استرجاع المعلومات من البيانات العربيّة غير المهيكلة

يرتكز نجاح شبكة الإنترنت أساساً على تكنولوجيا استرجاع المعلومات في محركات البحث مثل ياهو وجوجل وغيرها. ففي الآونة الأخيرة، جرى تفعيل نظم البحث الدلالي لاسترجاع المعلومات في العديد من التطبيقات الحياتية مثل الطب والتجارة الإلكترونية. وعلاوة على ذلك، تستخدم تلك النظم وكلاء البرمجيات (**software agent**)، التي تطبق مفاهيم الأنطولوجيا – علم التوصيف(**Ontology**) لتحديد معنى البيانات على مجال معين. فمن أهداف هذه الأنظمة أيضاً تحسين البحث عن طريق إزالة الغموض المرتبط بالاستعلام، وبالتالي استرجاع النتائج ذات الصلة. كما تعتمد الشبكة الدلالية على التقنيات والتكنولوجيا التي تسمح للآلات بقراءة وفهم واسترجاع معنى معلومات محددة من الإنترنت. وفيما يتعلق بمجال اللغة العربيّة، لم يُجرَ سوى القليل من البحوث على معالجة الشبكة الدلالية للبيانات غير المنظّمة. فضمن نطاق استرجاع المعلومات العربيّة، يتطلب نجاح الشبكة الدلالية وجود أدوات وتطبيقات داعمة مناسبة. وهكذا، فإنّ هذا العمل يعرض نظاماً جديداً باستخدام تقنيات معالجة اللغات الطبيعيّة وتقنيات الشبكة الدلالية لاسترجاع المعلومات وإدارة المعرفة المتعلقة بموارد البيانات العربيّة.

المسألة الرئيسة التي تواجه البحوث في مجتمع المعلومات حالياً هي تدفق المعلومات؛ وهذه مشكلة تفاقمت بسبب التنوع الهائل للمعلومات على الشبكة العنكبوتية العالميّة **World**

Wide Web))، وأعطت الباحثين إمكانية الوصول إلى الملايين من المراجع والمقالات والأخبار والخدمات. وبغض النظر عن الموقع الجغرافي واللغة المستخدمة، فإنّ الكثير من هذه المعلومات هي بيانات غير مهيكلة. وهناك مجموعة كبيرة من الأبحاث حول التنقيب في بيانات الويب (**Data Mining Web**) غير المهيكلة، ولكن لم يبذل إلا قليل من الجهد لصفحات الويب المؤلفة باللغة العربيّة.

تبحث هذه الدراسة في تحديات استرجاع المعلومات العربيّة كما هو مطبّق على كل من تقنيات برمجة اللغات الطبيعية (**NLP**) وتقنيات الشبكة الدلالية (**Semantic Web**) من خلال مراجعة الأعمال المنشورة وتجربة التقنيات الحالية. وقد ثبت أن استرجاع المعلومات العربيّة ليس بمستوى استرجاع المعلومات اللاتينية من حيث نضج تكنولوجيات المعالجة. وللتغلب على أوجه القصور، اقترح الباحثون نموذجاً هجيناً (**hybrid framework**) يتألف من تقنيات معالجة اللغات الطبيعية والتقنيات الدلالية. ويستخدم الإطار المقترح التقنيات الأساسية لمعالجة اللغات الطبيعية والقواعد النحويّة التي وُضع نموذجها بشكل مناسب بوضع علامة على النص العربي بشكل دقيق لنطاق معيّن، وسوف يتبعها استدلال المعرفة في قاعدة المعرفة الدلالية على غرار مناسب لاكتشاف حقائق جديدة في النص. كما أثبت الباحثون طريقة استخدام مجموعة بيانات دلالية عامة من مجموعة البيانات المرتبطة لحلّ أيّ غموض في البيانات التي يجري التنقيب عنها. سيركز عمل الباحثين في المستقبل على تحسين ذكاء قاعدة المعرفة الدلالية التي تقوم بنشر علاقات كيان متطورة (**sophisticated entity relationships**)، من شأنها أن تساعد في المستقبل في تصنيف المعرفة الدلالية لاستنتاج حقائق جديدة عن النطاق المستهدف.

International Journal on Islamic Applications in Computer Science And Technology, Vol. 5, Issue 1, March 2017,, 12-22

Year of publication: 2017

Evaluation Criteria for Computational Quran Search

Mohammad Alqahtani, Eric Atwell

سمة تقييم البحث الحاسوبي في القرآن

ملخص

يستعرض هذا المقال أدوات البحث المبنية من أجل استرجاع المعلومات من القرآن الكريم. ويقيم هذا المقال أدوات البحث المختلفة حسب ثلاثة عشر معياراً استناداً إلى: خصائص البحث وخصائص المخرجات ودقة الآيات المسترجعة واسترداد حجم قاعدة البيانات وأنواع محتويات قاعدة البيانات. بناءً على هذه الدراسة، نستنتج بأن معظم أدوات البحث الحالية في القرآن لا يمكنها حل مشكلة الغموض في النتائج المسترجعة، كون هذه الأدوات تستخدم تحليل الاستعلام التقليدي، كذلك استخدامها المحدود لأنطولوجيا القرآن الكريم.

المنهجية

يصف هذا القسم الإجراء المستخدم لتقييم أدوات البحث في القرآن، ويتحقق ذلك عن طريق اختيار معايير مشتركة لتقييم أدوات البحث المختلفة، ثم تقييم أدوات البحث الحالية في القرآن، وهي التي نوقشت قبل هذا البحث حسب هذه المعايير. الهدف الرئيس لهذا التقييم هو العثور على الأسباب الرئيسة للعوائق والقيود في أدوات البحث الحالية.

معايير مقارنة أساليب البحث في القرآن

تعتمد منهجية تقييم أساليب البحث في القرآن بالأساس على خوارزميات البحث ودقة النتائج وحجم قاعدة البيانات. ولُخصت المعايير الأكثر شيوعاً في مقالات عديدة تتعلق بمنهجيات تقييم محركات البحث. وتُعدّ المقاييسُ المشتركة خصائصَ البحث، على سبيل المثال معاملات بوليان (**Boolean operators**) والترتيبات ذات الصلة (**relevance rankings**) واسترداد (**recall**) النتائج ودقّتها (**precision**) وحجم قاعدة البيانات ووقت الاستجابة ونوع الاستفسار ومحتويات قاعدة البيانات. وقد أوضحت معايير التقييم المفصلة لأدوات البحث في القرآن بشكل جدول.

الخاتمة والعمل المستقبلي

يُلخّص هذا المقال تقنيات البحث في أدوات البحث الحالية في القرآن الكريم، مثل تطبيقات سطح المكتب وتطبيقات الإنترنت. كذلك يستعرض هذا المقال بحثا سابقا يتعلّق بأدوات البحث القرآنية، وقد جرى تقييم أدوات البحث القرآنية حسب ١٣ معياراً. بالاستناد إلى هذه الدراسة، عُثر على العديد من العيوب. أولاً، القيود المفروضة على أدوات البحث الحالية في القرآن الكريم لاسترجاع كافة المعلومات المطلوبة، حيث لا تدفع أدوات البحث المستخدمين إلى البحث عن المفاهيم أو العبارات أو الجمل أو الأسئلة أو المواضيع. هنالك أيضاً عيب آخر هو أنّ معظم أدوات البحث تستخدم موردًا واحدًا فقط، أو جزءا من نصوص القرآن الكريم، مما أثَّر على دقة النتائج المسترجعة. بالإضافة إلى أنّه لا توجد أداة بحث تتواءم مع كينونة القرآن الكريم لحل الغموض في النتيجة المسترجعة، هذه الأدوات لا تستخدم أساليب متقدمة في تحليل نصوص الاستعلام من خلال تطبيق تقنيات معالجة اللغات الطبيعيّة، على سبيل المثال التحليل والتدقيق الإملائي. هناك قيد نهائي على هذه الأدوات هو أنه لا توجد قوائم بالكيانات المسماة باللغة العربيّة المنسقة بشكل جيد وتختص بالنص القرآني. على سبيل المثال أسماء الأنبياء وأسماء الله الحسنى والحيوانات والأزمان والدين، بالتالي لا يمكن لأدوات البحث استخدام نظام التعرف على الكيانات المسماة (**Named Entities**). وسيطور العمل المستقبلي أساليب بحث دلالية تتضمن حلولاً لهذه القيود.

Artificial Intelligence Review, Volume 46, Issue 1, pp 113–128, June 2016

Year of publication: 2016

Hadith data mining and classification: a comparative analysis

Mohammad Arshi Saloot, Norisma Idris Rohana Mahmud, Salinah Ja'afar, Dirk Thorleuchter, Abdullah Gani

التنقيب في بيانات الحديث النبوي وتصنيفه: دراسة مقارنة

تُعدُّ الأحاديث النبوية مصدرا نصيّا مهمّا للتشريع والأعراف والتربية في العالم الإسلامي، ودراسة سمات الحديث اللغويّة الفريدة من نوعها (مثل اللغة العربيّة القديمة، والنصوص التي تتبع أسلوب القصة) تساعد على صياغة طرق خاصة لمعالجة اللغات الطبيعية والانتفاع بها. لكن فيما يختص بجانب الذكاء الاصطناعي، لا توجد دراسة تهتم بالحديث النبوي بشكل منفرد، وطرأ على هذا المجال العديد من المستجدات التي أُغفلت، ويتعيّن تسليط الضوء عليها. لذلك، يقوم هذا البحث بتحليل جميع المجلّات الأكاديمية والمنشورات الصادرة عن المؤتمرات باستخدام نظامين من أنظمة الذكاء الاصطناعي لدراسة الحديث النبوي، هما: التنقيب في بيانات الحديث، وتصنيفه. إنّ جميع الطرق والخوارزميات ذات الصلة بالحديث عُولجت وحُللت ضمن هذه الدراسة من ناحية الوظيفة، والسهولة، ومتوسط درجة الدقة والاستدعاء (score-F)، إضافة إلى معيار الدقة. و يجعل استخدام قواعد بيانات مختلفة ومتنوعة من الحديث النبوي عمليةَ المقارنة المباشرة بين نتائج التقييم أمرا مستحيلا، لذلك اعتمد البحث عادة على تطبيق الأساليب السابقة وتقييمها باستخدام نوع واحد من قواعد البيانات (٣١٥٠ حديثا من كتاب صحيح البخاري). إنّ نتيجة التقييم الذي أجري على طريقة التصنيف تظهر أنّ الشبكة العصبية قامت بإدراج الحديث النبوي ضمن نسبة دقة تقدر بـ ٩٤٪ ، ذلك لأن الشبكة العصبية قادرة على معالجة مدخلات على درجة عالية من التعقيد (ذات أبعاد كبيرة). وتحرز طريقة استخراج الحديث التي تجمع بين نموذج الفضاء الشعاعي (vector space model)، وتماثل جيب التمام

(Cosine similarity)، والاستعلامات المدعمة (enriched queries) أفضلَ نتائج من حيث الدقة من بين طرق التنقيب الأخرى التي أعيد تقييمها، وهو ما يصل إلى نسبة (٨٨٪). ويعد التوسّع في الاستعلام الجانب الأكثر أهمية في هذه الطرق، كما أنه يجب على هذه الاستعلامات أن تتناسب مع اللغة الخاصة بالحديث. وغياب الأساليب القائمة على المعرفة يعد واضحا في طرق التنقيب في الحديث وتصنيفه، ويمكن تغطية هذا الغياب في الأعمال المستقبلية باستخدام التمثيل البياني للمعلومات.

هنالك أنواع مختلفة من المعلومات التي يمكن استخراجها من الحديث النبوي، مثل نوع من أنواع المعرفة، بما في ذلك التشريعات الإسلامية (Islamic legislative)، ونظام الجيش الإسلامي (Islamic military)، وعملية تصنيف الحديث. لقد أُعد ١٣٢١ حديثا من كتاب صحيح البخاري لتجربة هذا النظام وتصويبه، مقسمة إلى أكثر من ثلاث عشرة مجموعة، هي: العقيدة، الفقه، الصلاة، الأذان، الخسوف، الزكاة، حسن الخلق، الصوم، التداوي، الطعام، الحج، المظالم، مناقب النبي محمد صلى الله عليه وسلم. وقد مرّت هذه التجربة بثلاث مراحل: في المرحلة الأولى طُبّقت المعالجة المبدئية التي تمثلت في إزالة الإسناد، وتقسيم البيانات، وإزالة كلّ من علامات الترقيم وعلامات التشكيل، والكلمات الشائعة، إضافة إلى عملية التجريد. بينما طُبّق في المرحلة الثانية القيام بعملية التصويب، بحيث بُنيت مصفوفة السمات (feature matrix) (المفردات والسمات) باستخدام طريقة (TF-IDF). أما المرحلة الثالثة فقد أتاحت القيام بعملية التصنيف، واستُخدمت مجموعة البيانات الناتجة عن عملية التصويب في المرحلة السابقة. بالإضافة إلى ذلك، فإنّ سمات الاستعلام تزيد من تأثير عملية المعالجة المنجزة في المرحلة الثالثة، وتساعد على التوسّع في هذا الاستعلام. أخيرا، جرى الوصول إلى هذا الصنف عن طر يق استخدام جدول يحوي على معامل التشابه (similarity coefficient)، بالإضافة إلى استخدام طريقة الحد الأقصى الإجمالي للتشابه (maximum cumulative similarity)، ووصلت نسبة الدقة إلى ٤٥٪ و نسبة ٤٩٪ لـ(F-score).

International Journal of Advanced Computer Science and Applications, 7(8), 223-230.

Year of publication: 2016

Ontology-based Query Expansion for Arabic Text Retrieval

Waseem Alromima, Ibrahim F. Moawad, Rania Elgohary, Mostafa Aref

توسيع الاستعلام القائم على الأنطولوجيا لاسترجاع النصوص العربيّة

تقوم هذه الورقة البحثية ببناء نظام استرجاع معلومات يرتكز على نطاق دلالي للمفاهيم والاستعلامات، وذلك بتوسيع الاستعلام اعتماداً على الأنطولوجيا (علم التوصيف) للمصطلحات. وقد تناولت العديد من الأبحاث هذا الموضوع، لكنّ القليل منها تطرّق لتطبيقه على نصوص القرآن الكريم. علاوة على ذلك، فإنّ الأبحاث المتوفرة في هذا المجال تعاني من عدم الفعالية نتيجة خصوصية اللغة العربيّة من حيث غموض التراكيب الموجودة فيها وتعقيدها وقواعدها.

لذا جاءت فكرة هذا البحث لتحسين فعالية الاستعلامات المسترجعة اعتماداً على نطاق التوصيف، وتفوقت الطريقة المقترحة على مثيلاتها من حيث الدقة (**precision**) والاستدعاء (**recall**) عند تجريبها، إضافة إلى إثراء هذه الورقة البحثية لعلوم الدين الإسلامي وباحثي اللغة وتطبيقات الويب الدلالي (**Semantic Web applications**).

اعتمد الباحثون في طريقتهم المقترحة على مرحلتين أساسيتين، إحداهما غير متصلة بالإنترنت (**Offline**)، وثانيهما عن طريق الاتصال بالإنترنت (**Online**)، حيث تكونت المرحلة الأولى من ثلاث خطوات رئيسية، هي: المعالجة القبلية للنصوص (**preprocessing of text**) التي تهدف إلى تحويل النصوص إلى مفردات (**tokenization**)، والتخلص من كلمات الربط الزائدة (**stop words**) مثل حروف الجر، ثم حساب أوزان الكلمات اعتمادا على تكرارها. وفي الخطوة الثانية

من المرحلة الأولى قام الباحثون بعملية الفهرسة (Indexing) للمفردات الناتجة باستخدام طريقة نموذج المتجه الفراغي (Vector Space Model) لتمثيل المفردات والاستعلامات على شكل متجهات، ثم في الخطوة الثالثة بُنيت وحدات الأنطولوجيا (Ontology module) باستخدام لغة الأنطولوجيا للمواقع (Web Ontology Language).

اعتمدت المرحلة الثانية على الاتصال بشبكة الإنترنت، وتضمنت أربع خطوات أساسية، بدأت بالخطوة الأولى المتعلقة بواجهة المستخدم (Interface-User) التي سهّلت عمليات إدخال الاستعلامات وعرض النتائج المسترجعة. أما الخطوة الثانية من هذه المرحلة فقد اهتمت بالمعالجة المسبقة للاستعلام المدخل، من خلال عمليات تحويل الاستعلام للمفردات وإزالة الكلمات الزائدة وغيرها، بعد ذلك أجريت عمليّة توسعة الاستعلام (Query Expansion) اعتمادا على نطاق الأنطولوجيا (Ontology Domain) الخاص به، وذلك عن طريق إيجاد العلاقات بين الكلمات، على سبيل المثال وصف « أماكن ما بعد الحياة» قد يعبر عن الجنة أو النار ومرادفاتها. أما في الخطوة الأخيرة من هذه المرحلة، فقد استُرجعت المعلومات بعد ضم الاستعلام الأصلي إلى الاستعلام بعد التوسعة. وللتحقق من فاعلية النظام المقترح أُجريت عملية المطابقة بين النصوص الناتجة من المرحلة الأولى مع الاستعلام الموسّع من المرحلة الثانية باستخدام طريقة مطابقة جيب التمام (cosine similarity).

Antoine Tabbone et Thierry Paquet. Colloque International Francophone sur l'Ecrit et le Document, Oct 2008, France. Groupe de Recherche en Communication Ecrite, pp.97-102.

Year of publication: 2008

A search engine for Arabic documents

T. Sari, A. Kefali

محرّك بحث مقترح للوثائق العربيّة القديمة (المخطوطات)

طُبّق النظام المقترح ضمن مرحلتين: الأولى تهدف إلى تمثيل كلّ وثيقة في قاعدة البيانات بكلماتها، بينما في المرحلة الثانية يمكن للمستخدمين البحث في المجموعة عن طريق كتابة بعض الكلمات الرئيسية في واجهة المستخدم التي صُمّمت لهذا الغرض. تضمنت المرحلة الأولى عدداً من الخطوات بدأت بالمعالجة القبلية (**preprocessing**) للنصوص عن طريق تحويل الصور لتدرج اللون الرمادي (**Gray level transformation**) والتحويل للنظام الثنائي (**Binarization**) والتنعيم (**Smoothing**). بعد المعالجة القبلية تُقسّم المخطوطة إلى سطور (**Line segmentation**) ، ويجري التخلص من السطور البيضاء ثم تُقسّم السطور إلى كلمات ومقاطع (**Sub word decomposition**) وضم العناصر المتشابهة (**connected component**). أما في مرحلة واجهة المستخدم، فقد ركز الباحثان على البحث عن مخطوطة عن طريق الكلمات المفتاحية.

تجدر الإشارة هنا إلى أن الباحثين استخدما خوارزمية مطابقة سلسلة تقريبية (**string-matching approximate**) للبحث عن رموز الملف استناداً إلى استخدام خوارزمية لحساب المسافة بين متجهات الكلمات سميت بـ «مسافة التحرير ليفينشتاين» (**Levenshtein edit distance**). وأثبتت نتائج الاختبار التي أجريت فاعلية التجزئة إلى الكلمات الفرعية في تمثيل المخطوطات اليدوية العربيّة، ونجاح طريقة الاسترجاع على مجموعة من الوثائق التاريخية.

ومع ذلك، لا تزال هناك بعض المشاكل التي لم تحل مثل تصحيح الوثيقة واستخراج الميزات من النصوص بفعالية كبيرة مما يتسبب في مشاكل في أداء الاسترجاع.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 145–151, Berlin, Germany, August 7-12, 2016.

Year of publication: 2016

A Personalized Markov Clustering and Deep Learning Approach for Arabic Text Categorization

Vasu Jindal

أسلوب تخصيص عنقدة ماركوف والتعليم العميق لتصنيف النصوص العربيّة

أصبح تصنيف النصوص مجالاً رئيسياً للبحث في مجتمع معالجة اللغات الطبيعية، وهناك تزايد كبير في عدد الأعمال في هذا المجال. ومع ذلك، فإنّ هذه الأعمال تركز على اللغة الإنجليزيّة وتتجاهل لغات سامية مثل العربيّة. هذا البحث يقترح تقنية جديدة تتألف من ثلاث مراحل لتصنيف الوثائق العربيّة إلى فئات مختلفة استنادا إلى الكلمات التي تحتوي عليها، حيث استُفيد من جذور الكلمات في اللغة العربيّة ودمجها مع مزيج من عنقدة ماركوف (Markov clustering) وشبكات المعتقد العميق (Deep Belief Networks) لتصنيف الكلمات العربيّة في عناقيد (clusters).

النموذج المقترح يتكوّن من ثلاث مراحل أساسية، المرحلة الأولى هي المعالجة المسبقة تليها خطوتان من التقسيم ثم في المرحلة الأخيرة التعلم. في مرحلة المعالجة الأولية تُقسَّم المستندات وتُجَزَّأ إلى كلمات متقطعة ، ثم يُستخرج الجذر وتُحسَب نسبته التكرارية أو وزن الجذر (root-word weighted) عن طريق معادلة شائعة تُسمّى (TFIDF)، وهذه المعادلة تحسب نسبة تكرار الجذر في مستند وعدم تكراره في بقية المستندات. في المرحلة التالية تُقسّم الوثائق من خلال مزيج من عنقدة ماركوف وأسلوب المعنى الضبابي. وفي الخطوة الأخيرة استُخدمت نماذج التعلم العميق وطُبّقت على كلّ مجموعة نتجت من الخطوة السابقة لتخصيص التعلم لكل مجموعة من الجذور العنقودية.

جرى تقييم الأسلوب المقترح باستخدام مجموعتين من البيانات الشائعة المستخدمة سابقا في تصنيف النصوص العربيّة، التي تحتوي على عشرة آلاف وثيقة مأخوذة من موقع الجزيرة الإلكتروني (Aljazeera) المتوفر على الرابط التالي : ((http://www.aljazeera.net، بالإضافة إلى ستة الأف وثيقة مأخوذة من وكالة الصحافة السعودية (Saudi Press Agency). عُرضت النتائج باستخدام أسلوب التحقق المعتمد على نظام الطيَّات (fold cross validation) وهو أسلوب مشهور في مقارنة مثل هذه القضايا، وجرى تطبيق عشر طيات تجريبية على النتائج لإجراء المقارنة. وحقق الأسلوب المقترح (نتائج دقة) وصلت إلى (٩١,٢٪،) واستدعاء وصل إلى (٩٠,٩٪) ونتائج مقياس «ف» (F-measure) وصلت إلى (٩١,٠٢٪). ومن الجدير بالذكر أنّ التقسيم طُبّق على مجموعات تتكون كل واحدة منها من (١٢٠٠٠) وثيقة، حيث أخذت عينات منها بشكل عشوائي من المصدرين المشار إليهما أعلاه.

٣-٤-٦ أبحاث الترجمة الآلية

تضمنت أبحاث الترجمة الآلية ستة أبحاث، بينها بحث تاريخ التّرجَمة الآلية وتَطوّرها: وهو دِراسَة مَسْحيّة للتّرجَمات العَربية-الإنجليزيّة، وبحثان نوع أ، هما نموذج هجين لدعم الترجمة الآلية الإحصائية المعجمية وبناء ذخيرة لغوية بتحرير لاحق للترجمة الآلية العربية: تعليقات وقواعد استرشادية ، وثلاثة أبحاث نوع ب، هي: الترجمة الآلية الهجينة المطبّقة على أنظمة التحكم بالوسائط، والترجمة الآلية العصبية بالكلمات والأحرف للغة العربية الحديثة، ونظام مقترح للترجمة الآلية الدلالية لترجمة النص العربي إلى لغة الإشارة العربية.

Current Journal of Applied Science and Technology, 23(4): 1-19, 2017; Article no.CJAST.36124

Year of publication: 2017

Machine-Translation History and Evolution: Survey for Arabic-English Translations

Nabeel T. Alsohybe1, Neama Abdulaziz Dahan and Fadl Mutaher Ba-Alwi

تاريخ التّرجَمة الآلية وتطوّرها: دراسة مسحيّة للتّرجمات العربية-الإنجليزيّة

المقدّمة

شهدت تكنولوجيا الاتصالات والمعلومات خلال العُقود الأخيرة نموّاً سريعاً، ممّا انعكس على نموّ قطاع الأعمال والأنظمة المعتمدة على الاتصالات التي يستطيع الأشخاص بواسطتها التّواصل مع غيرهم في المجتمعات المختلفة. لقد أصبحت التّرجمة اليوم نشاطاً تجارياً مهماً، خاصّة مع تسارع وتيرة النمو في تكنولوجيا المعلومات والاتّصالات التي تتطلب تطوير الآلات والأدوات التي تُسهّل عمليّة التّواصل بين الجانبين. لذلك، أصبَحت اللغة ركيزةً مُهمّةً في عقد الصّفقات مع أشخاص ينتمون لمجتمعات وثقافات مختلفة. ومع ازدياد الحاجة لفهم الرسائل المتبادلة بين الأفراد في سياق الأعمال التّجارية أو لأغراض الدراسة أو حتّى الصّداقة، أصبحت الحاجة مُلِحّة لوُجود طريقة كفُؤة للتَرجمة الآلية.

تُقدم هذه الدراسة المسحية ملخّصاً حول التّرجمة، ثم تناقش الأنواع المختلفة من التّرجمة الآلية. وأخيراً، تُناقش بعض الأدوات اللاّزمة في التّرجمة الآلية، ثمّ تختتم الدراسة بالاستنتاجات.

مشكلة البَحث

لسوء الحظ، تعدُّ التّرجمة مُهمّةً صعبة بسبب اعتمادها على معالجة اللغة الطّبيعية (NLP) والقواعد اللغوية. والطّريقة الوحيدة لتكون التّرجمة ذات معنى هي استخدام الذّكاء

الاصطناعي، لأنه يَدمِج مُختلف أنواع العلوم المطلوبة في عملية الترجمة. وقد ظهر حديثاً مجالٌ جديدٌ في إطار التعلّم الآلي وهو تطوير خوارزميّات لمعالجة اللغات الطّبيعية. وهذه الخوارزميات وظيفتها تحليل نصوص المصدر إلى أقسام الكلام الخاصة بها وإعطاؤها علامات (شارات) لتسهيل آليّة التّرجمة. وهي عبارة عن خوارزميات التنقيب عن البيانات (data mining) ، مثل أدوات أقسام الكلام التي تتخذ شكل أدوات التّشفير(encoder) أو فَكّ التّشفير (decoder) في التّرجمة الآلية العصبيّة. كما أن اندماج هذه الخوارزميات (خوارزميات التّنقيب عن البيانات) أو اتحادها مع أدوات أخرى مثل المحلّل الصّرفي أو المعرب اللغوي المستخدمة في التّرجمة الآلية الإحصائيّة، يساعد خلال عمليّة التّرجمة.

وهناك العديد من الدّراسات التي تقترح ترجمات بمساعدة آليّة، لكنّ التّرجمة الآليّة التي وهي حقل بحثي حديث، تعني التّرجمة المؤتْمتة بالكامل، وتختلف التّرجمة الآلية عن الترجمة بمساعدة الحاسوب أو التّرجمة البشرية بمساعدة آليّة أو التّرجمة الآلية بمساعدة بشريّة. وقد وصفت العديد من البحوث العملية الإلكترونية للترجمة الآلية، وجاءتْ هذه الدراسة لاستقصاء ما توصّلت إليه البحوث السّابقة ومناقشة الأدوات اللاّزمة للتّرجمة الآلية أملاً في توفير ملخّص حول الدراسات المتعلّقة بالتّرجمة الآلية للباحثين في هذا المجال في المستقبل.

المبحث الأول: التّرجمة

تشهد التّرجمة نمواً سريعاً جداً هذه الأيام، فهي وسيلة فعّالة لنقل الثقافة واللغة بسبب الاتصال مع الآخر. والجانب الرئيسي لها هو فهم معاني الكلمات في مفردات اللّغة، وأيّ نص مُترجم يُحكم عليه بالقبول من المراجعين أو المدقّقين عند قراءته كما هو مقصود بلُغةِ المَصدر، وبِعبارةٍ أخرى، لا تتمّ الترجمة حرفياً، وذلك لأنّ المؤلف قد يُعبر عن أفكاره ومشاعره أثناء الكتابة أو الكلام.

ووفقاً لمجموعة البنك الدولي، تُستخدم اللغة العربية القياسية الحديثة في كتابة الكتب والعقود والمقالات والشّعر، فضلاً عن كونها لغة القرآن الكريم. لكنه من الصعب جداً

ترجمة اللهجات المحلية، نظراً للحاجة الشديدة إلى معرفة العلاقات المعقدة بين اللغة العربية ولهجات كل بلد على حدة. أما التّرجمة الحَرْفيّة التي تعني تّرجمة كَلِمَةً بِكَلِمَه، فلا تعدُّ الخَيار الأمثل، خاصةً عندما لا تستطيع معاني الكلمات وحدها أن تُعطينا المعنى المقصود من الجُملة. ويعتمد الاستخدام اللغوي والمعجمي المنتظم لهذا النوع من النصوص على السّياق وعناصر المحتوى، ويُشير تحديداً إلى لغة المصدر أو إلى ثقافات لغة ثالثة (أي خلاف لغة المَصدر أو اللغة المستهدفة). وتتأثر الهياكل (التراكيب) القياسية للنّص في الكتب أو الصحف بالتقاليد السّائدة في حينه. وتتأثر الترجمة بما يتوقعه القارئ بلغة المصدر، مع الأخذ بالاعتبار تخمين ما لديه من معلومات ذات صلة بالموضوع واللهجة التي يستخدمها. وينطبق الأمر نفسه على اللغة المستهدفة بما يخص النقاط الثلاثة الأخيرة: توقع القارئ والمعلومات واللهجة. ويراعى كل ذلك حينما كان ذلك ممكناً بمعزل عن حرفية النص المصدر. وتتأثر وجهات نظر المترجم وتحيزاته بعدة عوامل قد تكون شخصية أو ذاتية، ومن ضمنها أسباب اجتماعية وثقافية، وقد تتضمن أيضاً «عامل الولاء للجماعة» التي ينتمي لها، وقد تعكس فرضيات المترجم الوطنية أو السياسية أو العرقية أو الدينية أو الطبقة الاجتماعية التي يتبع لها أو الجنس أو غيرها. وقسّم أحد الباحثين النُصوص المترجمة إلى خمس مصفوفات: الأولى هي النوع، وهو ما يتعلّق بالعوامل الأدبية والفلسفية والأفكار والدين. والمصفوفة الثانية هي المصفوفة الثقافية، وتتعلق بأسلوب التعبير والقصص الثقافية والأمثال. والمصفوفة الثالثة هي المصفوفة الدلالية، وترتبط بمعاني المرادفات، والتنسيق، والاستعارات الأصلية. ورابعاً: تأتي الأمور المرتبطة بمُستوى الرسومات والجمل والقواعد والنصوص القرآنية. وأخيراً: التباين أو التنوّع: ويرتبط باللّهجات، والنّغمات والسّجلات الصّوتية الاجتماعية.

من جهة أخرى، يرجع عدم تشابه التّرجمة مع النّص المستهدف الذي لا يُشبه تماماً النص المصدر لعدة أسباب، من ضمنها تراكيب اللغة المصدر أو محتواها وذلك بسبب القيود المفروضة وفقاً للفروق الدلالية والاصطلاحية بين لغة المصدر واللغة المستهدفة.

ويرى بعض الدارسين أنّ أبرز التّحديّات التي تواجه التّرجمة هي التنوع الكبير في

التراكيب أو القواعد اللغويّة وعملية تفسير الجملة التي تستخدم تركيباً لغوياً معيّناً، ومن ثمّ اختيار أفضل تركيب للجملة في اللغة المستهدفة. لقد أصبحت الترجمة نشاطاً تجارياً منذ عام ١٩٩٠، حيث طُوّرت أولى التّرجمات الآلية الإحصائية من شركة IBM وكانت تهدف إلى أتمتة المهام الأساسية للتّرجمة. وهناك العديد من أنواع الترجمات التي تستخدم التقنية (التكنولوجيا) للمساعدة في الترجمة، وهو ما سيُناقش في المبحث التالي.

المبحث الثاني: التّرجمة الآلية

تُعدّ الترجمة الآلية إحدى أهم الحُقول وأصعبها من حيث إمكانية التطبيق في حياتنا اليومية، وهناك العديد من أنواع البحوث حول بناء التّرجمة الآلية أو تطويرها باللغة الإنجليزية وغيرها من اللغات. بالمقابل، هناك القليل من البحوث المماثلة حول اللغة العربية، بل تُعدّ مثل هذه البحوث غير موجودة.

على سبيل المثال، كانت الأوراق البحثية السابقة ذات الصلة باللغة العربية (و لغات أخرى مثل اللغة العبريّة) عبارة عن ترجمة بشرية بمساعدة آلية، أو مجرد أبحاث حول معالجة اللغة الطّبيعية لغرض تطوير محلّل أو مُعرب لغوي يحتوي على العديد من أخطاء التّرجمة. في حين طورت الأعمال البَحثيّة الحديثة ترجمة آليّة عصبية معتمدة على التّرجمة الآلية الإحصائيّة (Neural Statistical MT)؛ التي يمكن قياس أدائِها بالمقارنة مع التّرجمة الآلية الإحصائية (Statistical MT)، وباتت غير مقنعة بسبب الكمّ الهائل من الأخطاء التي تنطوي عليها النصوص المترجمة بواسطتها.

توفّر لنا الشّبكة الدّلالية وسيلة لتطوير التّرجمة الآلية، ويمكن أن يكون أداؤها أكثر إقناعاً مقارنة بأداء التّرجمة الآلية الإحصائيّة، وذلك وفقاً للنتائج التي عرضها الباحثون الذين قاموا بتطوير التّرجمة الآلية المستندة إلى تقنيات الشّبكة الدّلالية.

التّرجمة الآلية الإحصائيّة

"الترجمة الآلية الإحصائية" هو مصطلح يشير إلى مجموعة من أنظمة الترجمة الآلية التي تم تطويرها باستخدام أساليب التعلّم الآلي، وهي النوع الأكثر دراسة من أنواع الترجمة الآلية. لقد طُورت أولى أنظمة الترجمة الآلية قبل أقلَّ من ثلاثة عُقود، واستخدمت شركة IBM في عام ١٩٩٠ نظرية Bayes لتطوير منهجية إحصائية للتّرجمة الآلية، حيث اعتمدت على اتخاذ سلسلة من الكلمات والرموز الموجودة في لغة المصدر مع مجموعة من المفردات يمكن أن تُسمّى (أ) وتحويلها إلى سلسلة من الكلمات والرموز الموجودة في اللغة المستهدفة مع مجموعة أخرى من المفردات تسمى (ب).

التّرجمة الآلية باستخدام الشّبكة الدلاليّة

تعمل الشّبكة الدلالية على إعادة هيكلة كمّ هائل من البيانات التي يمكن الوصول إليها وتكون مفهومة لكل من البشر والآلات المتاحة على الشّابكة بطريقة مشابهة لتلك التي يُدركها العقل البشري، وتكون بمثابة تدريب للشبكة على فهم السياق القريب من أي كلمة أو عبارة يتمّ البحث عنها.

التّرجمة الآليّة العصبيّة

تُعدّ التّرجمة الآلية العصبية موضوعاً جديداً، بدأ يشهد في الآونة الأخيرة نشاطاً، حيثُ تجري خلالها التّرجمة الآلية بشكل مختلف تماماً عن الطرق التقليدية لأساليب التّرجمة الآلية الإحصائية القائمة على العبارة. فبدلاً من التَمرُّن على المكوّنات المختلفة للترجمة الآلية بشكل مستقل، يستخدم هذا النموذج الشبكة العصبية الاصطناعية لتعليم النموذج أو تدريبه على جمع المكونات معا لتحقيق أكبر قدر من الإتقان في أداء الترجمة بواسطة خطوتي الشبكة العصبية المتعاقبتين: «التشفير وفكّ التشفير».

المبحث الثالث: بعض الأدوات اللاّزمة

يعرض هذا الجزء من الدراسة الأوراق البحثية المختلفة التي تُغطّي المكونات التي سيتم

استخدامها في النموذج المقترح، فيناقش هذا المبحث المحلّلات اللغويّة العربية، والمحللات الصّرفية، والمحلّلات الدّلالية بالإضافة إلى خوارزميّات التّصحيح.

أولاً: المحللات اللغوية العربية (Arabic Parsers)

كما أسلفنا أعلاه، تتضمن اللغة العربية تراكيب مختلفة، فالكلمة أو المفردة العربية تتكون من مقطعين: الجذر؛ وهو عبارة عن مجموعة الأصول، والنّمط؛ وهو عبارة عن مجموعة من حُروف العلّة، وهما يُشكّلان معاً جَذع الكلمة. بالإضافة لما سبق، لا يمكن للكلمات أن تشكل وحدها الجمل على هيئتها تلك (الجذع أو أصل الكلمة الاشتقاقي)، بل هناك حاجة لحروف إضافية أخرى من الممكن أن تُضيف معنى آخر للكلمة.

إنّ مُحلل اللغة الطبيعية هو آلة يمكنها أن تفهم أجزاء الجملة وتساعدنا في الترجمة باستخدام الترجمة الآلية، وقد قام العديد من الباحثين بدراسة هذا النوع من المحلّلات. وناقش هذا المبحث بعض أحدث أنواع المحلّلات العربية التي استطاعت أن تُحقّق نقلة نوعية في مجال الترجمة العربية ومعالجة اللغة الطّبيعية (ومن الأمثلة عليها تلك المحلّلات التي قامت بتطويرها شركة ميكروسوفت).

ثانياً: المحللات الصّرفية(Morphological Analyser)

الصّرف هو الدراسة والتحديد والتحليل والوصف لوحدات الحدّ الأدنى (المقطع الصّرفي) التي تَحمل معنى وتشكّل كلمةً واحدة. ويُعدّ الغموض الصّرفي مصدر قلق بالنّسبة للمحللات النّحوية، وغيرها من أدوات معالجة اللغات الطبيعية. ويعطي التّحليل الصّرفي معلومات أدق حول أجزاء الكلام بحيث يختار التّحليل الأنسب لها بشكل يتوافق مع السّياق، وهو وفقاً لبعض الباحثين «أحد مكونات المحلل اللغوي». في حين ذكرت إحدى الدّراسات أن « المحلل اللغوي هو أداة تحليل صرفي».

وأشارت إحدى الدراسات التي شملها هذا المسح إلى امتلاك اللغة الإنجليزية والبرتغالية نفس التركيب اللغوي، وبالتالي لم تكن هناك حاجة إلى استخدام مُحلِلَيْنِ اثنين مُختلفَيْن لمعالجة اللغة الطبيعية. وحسب الباحث، فقد كان التحليل الصّرفي كافياً لإجراء التحليل بتقسيم

الجملة إلى أجزائها الرئيسية (إما اسم أو فعل).

ثالثاً: المُحلّلات الدّلالية

المسألة الأساسية التي يجب أن تُعالِجها أنظمة التّرجمة الآلية هي الغموض، ويمكن القول بأن التّرجمة الآلية هي نقل لدلالات النّص من لغة الإدخال إلى لغة الإخراج. لذلك، لكي يكون هناك أنظمة متخصّصة، يجب تسجيل المعرفة ومعالجتها وتخزينها واسترجاعها وإعادة استخدامها ونشرها. وتُعدّ الأنطولوجيا الطريقة الملائمة للقيام بكلّ ذلك، ومن هنا جاء مصطلح الدّلالة. إذاً، لغة التّوصيف هذه هي «مواصفات صريحة للمفاهيم». وهي «تحدّد المصطلحات وما تحمله من علاقات محددة بينها، ويمكن تفسيرها من البشر والحواسيب». وذكر بعض الباحثين- في هذا السياق - وجود اختلافات بين اللغتين العربية والإنجليزية من حيث آلية عمل الأنطولوجيا، واقترحوا تعزيز تلك الآليات المتاحة لتشمل العربية.

رابعاً: إعادة ترتيب الجملة (خوارزميات التصحيح)

يجب أن تضمن التّرجمة الآلية دقّة ترجمة اللغة المستهدفة، سواء من الإنجليزية إلى العربية أو من العربية إلى الإنجليزية. لذلك، في كلتا الحالتين، يجب التحقق من تركيب الجملة. أشارت إحدى الدراسات إلى إمكانية التعامل مع الجمل في البُرتغالية والإنجليزية كشيءٍ واحد من حيث التّركيب. وذكرت كذلك بأنه لا يلزم إعادة ترتيب الجملة قبل صدور النصّ النهائي للغة المستهدفة لأن الجملة سوف تترتّب فعلياً بعد التّرجمة الإحصائيّة. ووفقا لذلك، فإن الاستبدال التلقائي لن يسمح بظهور أيّ خطأ. على النّقيض من ذلك، تتضمن اللغة العربية تراكيب مختلفة، بحيث يحل الاستبدال التلقائي محل الجناس اللفظي وترجمته النسبية وفقاً للسياق.

الاستنتاجات

أصبحت الترجمة النشاط الرئيسي في حياتنا اليومية بسبب أهميتها في الأنشطة التقليدية والدراسة والمحادثات بين الأصدقاء، بل وأحياناً في العقود أو الاتّفاقات السّياسية والقانونية. والترجمة لها العديد من الجوانب والأنواع والسّمات. وباتت التّرجمة مؤخراً نشاطاً تجارياً، خاصة

في ظل تسارع وتيرة النمو في قطاع تكنولوجيا المعلومات والاتصالات، وهذا يتطلب تطوير الآلات والأدوات التي تُسهّل التواصل بين الأطراف المختلفة.

والتّرجمة الآلية هي واحدة من تلك الأدوات اللاّزمة؛ بل تُعدّ الجندي المجهول في أيّ عملية من عمليات الاتصال. ومع ذلك، لم تُفلح التّرجمة الآلية -حتى اللّحظة- في التفوق على التّرجمة البشرية. لكن لحسن الحظ، لم يتوقف البحث العلمي في هذا المجال، ونأمل أن تستمر مساهمته فيه. وهناك ثلاثة أنواع من الترجمة الآلية: التّرجمة الآلية الإحصائية والتّرجمة الآلية العصبية والتّرجمة الآلية باستخدام الشبكة الدلالية.

وفي بعض الأحيان، قد تكون هناك حاجة إلى بعض الأدوات الأخرى لاستخدامها في التّرجمة الآلية، للمساعدة في تحليل النص ومن ثم تطوير النصوص في اللغات المستهدفة. ولا تزال هذه البحوث مستمرّة ولن تتوقف أبداً حتى تصبح التّرجمة الآلية مكافئة أو أفضل من ترجمة الإنسان.

IJCSI International Journal of Computer Science Issues, Volume 12, Issue 2, March 2015

Year of publication: 2015

A Hybrid Model for Enhancing Lexical Statistical (Machine Translation (SMT

Ahmed G. M. ElSayed, Ahmed S. Salama, Alaa El-Din M. El-Ghazali

نموذج هجين لدعم الترجمة الآلية الإحصائية المعجمية

ازداد الاهتمام بالترجمة الآلية المستندة إلى الإحصاء بشكل كبير نتيجة الأحداث السياسية والاجتماعية في العالم. يقدم هذا البحث الترجمة الآلية الإحصائية مستندة إلى نموذج (model) يمكن استعماله في ترجمة الجمل من اللغة المصدر (الإنكليزية) إلى اللغة الهدف (العربية) بشكل ذاتي، بواسطة نماذج إحصائية خاصة بمعالجة اللغات الطبيعية مثل نموذج المحاذاة (alignment) ونموذج العبارات ونموذج إعادة التسلسل (reordering) ونموذج الترجمة. هذه النماذج جمعت سوية لغرض تعزيز أداء الترجمة الآلية الإحصائية، واستعملت أدوات عديدة في هذا البحث سنرد على ذكرها فيما بعد.

بالاستناد إلى تطبيق هذا النموذج وتقييمه قورنت الترجمة الناتجة مع طرائق أخرى للترجمة الآلية مثل جوجل، وظهر أن النتائج تعزز الطريقة الإحصائية وتثبت أنها طريقة كفؤة في هذا الحقل من البحث العلمي.

استعمل هذا النموذج عدة نماذج بمستويات متعددة:

نموذج اللغة، ونموذج التسلسل، والنموذج المعتمد على العبارات، ونموذج إعادة التسلسل، ونموذج التوليف (tuning model). واستعمل نموذج إعادة التسلسل من أجل معرفة الكلمات التي تقابل بعضها بين اللغتين المترجم منها والمترجم إليها، بحيث تكونان بالمعنى نفسه بين اللغتين.

مدير النظام

هو أول مكوّن من النموذج، ويستعمل واجهة المستخدم (user interface) والجمل التي سبقت ترجمتها في قواعد البيانات لاستعمالها أساسا للترجمة. كما يحتوي البرنامج على واجهة للمستخدم، وعلى قواعد البيانات ونظام إدارة قواعد البيانات، وعلى محضّر البيانات (Data Preparator) وهي التي تقوم بتحويل الذخيرة اللغوية المتوازية إلى صيغة تقوم بالمهام الأربعة التالية:

١- عملية التطبيع (Normalization) بتحويل كل الكلمات من المصدر إلى الهدف.

٢- العملية الثانية هي Tokenization للبيانات في كلا ذخيرتي اللغتين وذلك بوضع فراغات وترميز (punctuation) للكلمات.

٣- إضافة احتمالية لكل الكلمات في الذخيرة اللغوية المتوازية وبناء النموذج لذلك بهدف توليد الملف لكل لغة.

٤- تنظيف البيانات بحذف الجمل الطويلة التي هي أطول من حد معين وحذف الجمل الفارغة أو تلك التي ليست مرتبة ويمكن أن تؤثر في نوعية الترجمة.

المدرّب trainer

هذا هو قلب النموذج. يحتوي هذا المكوّن على عدّة نماذج لتنفيذ النموذج الهجيني لبيانات التدريب للحصول على نتائج ترجمة جيدة. يحتوي المدرّب على الأجزاء الآتية:

نموذج اللغة الذي يبنى للتأكد من أن الترجمة قد تمت بشكل سليم وطلاقة في اللغة المترجم إليها. ويستند هذا النموذج على ما يسمى نموذج ""ngram . هذا النموذج هو نموذج إحصائي يصف اللغة ويحتوي على ترددات:

token-based n-grams التي تتكرر في الذخيرة اللغوية. يجري تدريب نموذج اللغة من ذخيرة لغوية ضخمة بلغة واحدة (monolingual corpus) ويحفظ في ملف خاص. وهذا الملف يعتمد على المزج بين النموذج الأحادي unigram language model وبين النموذج

الخماسي gram language model٥.

نموذج المحاذاة (Alignment Model)

يستعمل هذا النموذج في الترجمة الإحصائية لمعرفة العلاقة بين الكلمات المترجمة من جملة في اللغة المصدر بالمقارنة مع الكلمات في الجملة التي تعطي المعنى نفسه في اللغة الهدف. إن عملية المحاذاة هذه تمثّل واحدة من أهم العمليات في عملية الترجمة التي تستعمل لإنتاج النص للكلمات المقابلة مع بعضها.

في عملية المحاذاة هذه يجري تعميم وسائل استخلاص العبارات من المنظومة المعتمدة على العبارات مع ما يقابلها في اللغة الأخرى. وتستعمل عملية المحاذاة لمقابلة الكلمات مع بعضها لتحسين الترجمة. البيانات المتحاذية هي عناصر الذخيرة اللغوية المتوازية بين اللغتين، وكل عنصر في إحدى اللغتين يوائم العنصر المقابل في اللغة الثانية، ويحتوي النموذج على النموذج التدريبي لمراقبة المحاذاة بين البيانات ويساعد على نموذج إعادة الترتيب.

مستخلص الترجمة والعبارات:

يستعمل هذا المكوّن للحصول على ترجمة مباشرة للكلمات، ومن ثم تكوين جدول لأفضل كلمة مناسبة من بين الخيارات المطروحة، ثم يستخلص جدول العبارات التي تصف الوضع الإحصائي للذخيرة اللغوية المتوازية بين الجمل المتقابلة بين اللغتين.

نموذج إعادة التسلسل

بعد استخلاص العبارات مع درجات أفضليتها، يقوم هذا المكوّن بإعادة سَلسَلة نموذج المحاذاة للحصول على جدول إعادة التسلسل. يحتوي هذا الجدول على الترددات الإحصائية التي تصف الكلمات بين اللغتين المصدر والهدف.

نموذج الترجمة

هذا هو المكوّن الأخير الذي يجب توليده باعتباره ناتجا أخيرا بعد تدريب كلّ المكونات السابقة التي يجب توليفها فيما بعد أو تستعمل لتوليد الترجمة مباشرة. إن نموذج الترجمة يحتوي على جدول ترجمة ونموذج محاذاة وجدول عبارات وجدول إعادة تسلسل ونموذج اللغة الذي يستعمل في الترجمة.

مكون التوليف

الهدف الرئيسي من هذا المكون هو دعم الترجمة الآلية الإحصائية الناتجة وتحسين نوعيتها. والنموذج الخطّي للتوليف يهدف إلى إيجاد أفضل أوزان تجعل الترجمة الناتجة أفضل ما يمكن، لذلك فإنّ عملية التوليف تهدف لإيجاد أفضل إعدادات لنموذج الترجمة. إنّ عملية التوليف تقوم بترجمة آلاف من الجمل من اللغة المصدر في مجموعة التوليف باستعمال نموذج الترجمة وتقارن نماذج الناتج مع مجموعات مرجعية مترجمة من مترجمين من البشر قاموا بذلك سابقًا، ثم بعد ذلك تُغيّر الإعدادات بهدف تحسين نوعية الترجمة إلى أن نصل إلى أفضل نوعية للترجمة مع تقليل نسبة الخطأ لأقل حد ممكن.

المترجم

هذا هو المكون الأخير في النموذج، ويُطبّق نموذج الترجمة بمكوناته (نموذج اللغة وجدول الترجمة ونموذج المحاذاة وجدول العبارات وجدول إعادة التسلسل) على الجملة المجّهزة من المستخدم لكي تترجم بشكل صحيح. إن مهمة هذا المكوّن هي تطبيق نموذج الترجمة على الجملة المطلوب ترجمتها وإعادتها إلى مرسلها.

نموذج التدريب وخوارزمية التعليم

يدعم مدير النظام المعجم ويغذّي بياناته بالنصوص والجمل المترجمة بين اللغتين في الذخيرة ثنائية اللغة. ويقوم بإدخال هذه البيانات من خلال واجهة المستخدم وأدوات نظام إدارة قواعد البيانات، ويعمل هذا النظام على:

تجهيز الذخيرة اللغوية المتوازية لعملية التدريب بواسطة كلمات التطبيع (normalizing words) وعمليات tokenization و true casing وعملية التنظيف.

- تكوين نموذج اللغة للغة المصدر للمساعدة في عملية الترجمة.
- تدريب نموذج محاذاة الكلمات باستعمال بيانات الذخيرة اللغوية المتوازية.
- حفظ ملفات المحاذاة للرجوع إليها ولكي تساعد على إيجاد جدول إعادة التسلسل.
- توليد جدول الترجمة بحيث يكون أكثر احتمالية من ناحية درجات الأفضلية.
- حفظ جدول الترجمة.
- بناء جدول العبارات للعبارات المترجمة مع درجات أفضليتها.
- حفظ جدول العبارات.
- بناء جدول إعادة التسلسل.
- حفظ جدول إعادة التسلسل.
- بناء جدول الترجمة.
- حفظ جدول الترجمة.
- توليف نموذج الترجمة وتقليل نسبة الأخطاء لتحسين نوعية الترجمة.
- تدقيق جدول الترجمة وجدول إعادة التسلسل من الذخيرة الثنائية المتوازية.

المقترحات المستقبلية

إنّ مسألة الترجمة الآلية لم تحل لحد الآن، وهناك حاجة للكثير من البحث العلمي في هذا المجال لكسب الطمأنينة للوصول إلى مستوى من الترجمة يضاهي الترجمة من مترجمين من البشر. إنّ الحاجة إلى ترجمة أسرع وأرخص بين اللغات ستتعزز فقط بمشاركة المعلومات بين الأمم. واستنادًا إلى ذلك نقترح ما يأتي:

إن عملية العنقدة (clustering) هي إحدى الوسائل المهمة التي يجب عملها باللغة العربية

لحل مشكلة **tokenization problem**. وهذه العملية هي طريقة عامة للتعامل مع التشتت والتحليل الصرفي (**morphological analysis**).

هناك حاجة لاستعمال كمية كبيرة من الذخيرة اللغوية (**corpus**) المتوازية (بلغتين) بهدف تدريب النموذج للحصول على نتائج أفضل.

يجب استعمال كمية كبيرة من البيانات من الحيز (**domain**) نفسه الذي تقع فيه بيانات التدريب للحصول على نتائج أفضل.

كلما قمنا بتنعيم (**tuning**) النموذج بشكل أفضل كانت النتائج أفضل، لذلك يجب القيام بعملية التنعيم قدر الاستطاعة.

يجب استعمال بيانات الفحص في الحيز نفسه للحصول على نتائج أفضل.

التكامل مع وسائل أخرى للترجمة الآلية مثل الشبكات العصبية الذكية (**artificial networks neural**) في عملية التعلّم وعملية التدريب للحصول على نتائج أفضل.

International Conference on Language Resources and Evaluation (LREC 2016)

Year of publication: 2016

Building an Arabic Machine Translation Post-Edited Corpus: Guidelines and Annotation

Wajdi Zaghouani, Nizar Habashy, Ossama Obeid, Behrang Mohitz, Houda Bouamor and Kemal Oflazer

بناء ذخيرة لغوية بتحرير لاحق للترجمة الآلية العربية: تعليقات وقواعد إرشاديّة

في هذا البحث نستعرض القواعد الإرشادية مع أسلوب التعليقات (annotation procedure) لتكوين ذخيرة لغوية يجري عليها تصحيح لاحق من البشر خاصة باللغة العربية القياسية الحديثة. وبالطبع فإنّ تكوين أية ذخيرة لغوية يدويًا يلاقي صعوبات بالغة. ولغرض توضيح هذه الصعوبات فقد كونّا تعليقات إرشادية مبسّطة استعملت من فريق عمل مكوّن من أشخاص قاموا بوضع هذه التعليقات. وبهدف التأكّد من توافق هذه التعليقات بين الأشخاص المختلفين فقد أُجري تدريب لهم وعُقدت جلسات مناقشة فيما بينهم. وحسب علمنا فهذه هي أول ذخيرة لغوية عربية يجري تحريرها وإجراء التعليقات عليها لخدمة الترجمة الآلية.

إنّ هذه الذخيرة هي جزء من بنك قطر للغة العربية، وهو مشروع ضخم لعمل التعليقات اليدوية. وهدف هذا المشروع هو تكوين ذخيرة تحوي مليوني كلمة للمستخدمين عبر الشابكة لجمع ملاحظاتهم على مواقع الأخبار وغيرها. وقد اختير في هذه المرحلة جزء مقداره مائة ألف كلمة بهدف الترجمة الآلية من مختلف المواقع الإخبارية الإنكليزية المترجمة للعربية باستعمال ترجمة جوجل كمرحلة أولى.

وصف الذخيرة

لقد جمعنا من عدد من المواقع الإخبارية على الشابكة مائة ألف كلمة كذخيرة من مقالات

الأخبار الإنجليزية. وبما أن الحصول على هذه الأخبار مجاني فقد تغلبنا بذلك على مشكلة حقوق النشر. لقد احتوت هذه الذخيرة على ٥٢٠ مقالة بمعدل ١٩٢ كلمة للمقالة، تغطّي معظم المقالات أحداثًا سياسيّة، وقد أُخذت من المقالات الأصليّة قبل أي تحرير عليها. وكانت الملفات الأصلية بصيغة HTML ثم حولت إلى صيغةUTF-٨ التي هي بصيغة نصوص خالصة قياسيّة لكي تكون قابلة للاستخدام لاحقا. وقد تُرجمت الذخيرة المجمّعة آليا من الإنجليزية إلى العربية باستعمال مترجم جوجل مدفوع الثمن.

كانت الخطوات الإرشادية لوضع الأخطاء التي يجب تصحيحها كالآتي:

الأخطاء الإملائية، وبالأخص أخطاء حرف الياء والهمزة.

أخطاء اختيار الكلمة، وهذه كثيرة الورود في الترجمة الآلية.

الأخطاء الصرفية، وغالبا ما تتعلق بالاشتقاق.

الأخطاء المتعلقة بالتذكير والتأنيث، وتوافق العدد، والتعريف والتنكير، وزمن الفعل.

الأخطاء المتعلقة بأسماء الأعلام عند ترجمتها إلى العربية.

استعمال اللهجات، فالترجمة الآلية لا تتعامل مع اللهجات المحلية.

الأخطاء في علامات الترقيم حينما توضع في غير مكانها.

إنّ دقة الترجمة مهمة جدًا، وأيّ نقص في الترجمة يجب إضافته للتأكّد من شمول كلّ الدلالات الواردة في النص الأصلي. وفي كلّ الأحوال فإنّ الترجمة الآلية بعد تنقيحها يجب التأكّد من أن تكون مطابقة للمعنى في الأصل الإنجليزي. ولا يُسمح بإعادة تسلسل الكلمات والعبارات في الجمل الطويلة إلاّ إذا لزم الأمر، حيث يُشجع الشخص الذي يضيف التعليقات ليستعمل قدر الإمكان نتائج الترجمة الآلية الخام (raw MT output). بعد ذلك فإن الترجمة الآلية التي نُقّحت يجب أن لا تضيف بأيّ حال من الأحوال أية معلومة إضافية، ولا تحذف أية معلومة أو معنى كان موجودًا باللغة الإنجليزية.

وفي حال وجود أخطاء في القواعد أو في التوافق في نصوص الترجمة الآلية، يجب تصحيح

ذلك دائمًا. وبما أن ناتج الترجمة الآلية يُولّد ذاتيًا، فإنّ الترجمة لها نسق خاص يبدو غير مألوف أو غير سلس، رغم أنه في بعض الأحوال يحوي كلمات بتسلسل مقبول وتعطي المعنى المراد بالنص الإنجليزي بدقة. وفي مثل هذه الحال فإنّ نسق الكتابة يجب ألا يتمّ تغييره أو تحسينه.

فمثلا لترجمة الجملة: **It's been five years since pro-democracy protests started»"**

كانت الترجمة الآلية: «إنها كانت خمس سنوات منذ بدأت الاحتجاجات المؤيدة للديمقراطية»

بينما بعد التنقيح تصبح: «لقد مرت خمس سنوات منذ بدأت الاحتجاجات المؤيدة للديمقراطية»

احتوى فريق التعليقات على خمس أشخاص، وكان رئيس فريق المعلّقين هو نفسه مدير عمل المشروع. وبذلك فقد قام بتقييم جودة عملية التعليقات ومراقبة تقرير تقدم العمل في التعليقات وكتابته.

لقد طُوّر بروتوكول (**protocol**) معرّف جيدًا يتضمن روتين (**routine**) لعملية التعليقات لما بعد الترجمة الآلية بحيث يقوم بتوزيع المهام (**job assignment**) وعملية تقييم التعليقات الداخليّة. (**Inter-annotator agreement evaluation**)

وقد كان رئيس فريق العمل مسؤولا عن اختيار الذخيرة وعملية التطبيع (**normalization**) إضافة إلى التعليقات على البيانات القياسية الذهبية (**gold standard data**) المستعملة لحساب توافق التعليقات الداخلية (**Inter-Annotator Agreement**) من ضمن الذخيرة.

تقدّم الواجهة الأنواع التالية من التصحيحات:

تحرير الكلمات، وذلك بتصحيح الكلمات أو تحويرها.

تحريك الكلمات إلى اليمين أو الشمال.

إضافة كلمة ضمن النص.

حذف كلمة لا لزوم لها.

دمج الكلمات وتفريقها.

توافق التعليقات الداخلية

استُعملت نسبة خطأ الكلمات (Word error Rate) دليلا على توافق التعليقات الداخلية، فإذا كانت نسبة خطأ الكلمات بين تعليقين مختلفين للجملة نفسها قليلة، عندها يُفترض أنّ هناك توافقا كبيرا بينهما. ولغرض قياس جودة ما بعد تحرير الترجمة الآلية، وجب قياس توافق التعليقات الداخلية على ملفات تُؤخذ بشكل عشوائي للتأكّد من أنّ التعليقات كانت متناسقة في اتباع تعليمات التعليقات. إنّ التوافق العالي في التعليقات هو مؤشر جيد على جودة البيانات. وقد أظهرت النتائج أن مقدار التغييرات في النص، أي ما استبدل كان بنسبة ٣١,٧٥٪ من النص.

تحليل الأخطاء

لا بدّ أن تكون هناك دائما حالات من عدم التوافق في تحرير ما بعد الترجمة الآلية، لكن هناك دائمًا طرائق متعددة لتصحيح ترجمة ما. لقد حاولنا تقليص عدم التناسق قدر الإمكان في التعليقات.

الخلاصة

لقد عرضنا طريقة استُعملت لترجمة مائة ألف كلمة من الإنجليزية إلى العربية، وقد أجريت عملية التحرير عليها يدويا وفق منهاج للتعليقات ومراقبة التحكم بالجودة بواسطة قياس التعليقات الداخلية. وُضِعت خطة إرشادية وستُعلن هذه الخطة، ونأمل أن نتعاون في عملية تحرير ما بعد الترجمة الآلية لتصحيح الأخطاء آليا ومن ثم الحصول على تغذية راجعة من مجتمع الباحثين لتقييم مدى فائدة هذه الطريقة وفعاليتها. نحن نعتقد أن هذه الذخيرة ستكون مفيدة في تقدّم الأبحاث في جهود الترجمة الآلية، لأننا نحتاج عادة للتحرير اليدوي للترجمة

الآلية. نحن نرى أن طريقتنا في وضع أسس إرشادية للتعليقات بطريقة متناسقة يمكن تطبيقها في مشاريع أخرى وفي لغات أخرى كذلك. وسنحاول في المستقبل زيادة حجم الذخيرة وأن تتضمن حقولا معرفية أخرى.

Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas, Waikiki, Hawai'i, USA, 21-25 October 2008

Year of publication: 2008

Hybrid Machine Translation Applied to Media Monitoring

Hassan Sawaf, Braddock Gaskill, Michael Veronis

الترجمة الآلية الهجينة المطبّقة على أنظمة التحكم بالوسائط

نقدّم في هذا البحث نظاما لتمييز الكلام المحكي باللغة العربية العاميّة من خلال تحويله إلى اللغة العربية المعاصرة لتتم ترجمته بعد ذلك إلى الإنجليزية ، وسُجّلت المدخلات من قناة تلفزيونية. من المتعارف عليه أن سماع الكلام المحكي وتحويله إلى نص مكتوب ثم ترجمته هما عمليتان منفصلتان، وعندما يكون هناك خطأ في تحويل الكلام إلى نص فستكون الترجمة خاطئة.

تمكّنت الترجمة الآلية من حلّ الإشكال في بعض الأحوال باستخدام معلومات إضافية مثل الترجمة الآلية المعتمدة على القواعد التي تساهم في تصحيح الأخطاء باستخدام معلومات دلالية، بينما تستخدم أنظمة الترجمة الآلية الإحصائية آليات فكّ تشفير معتمدة على الرسم البياني، وكلا النظامين له نقاط قوة وضعف، وقد ساهما بشكل منفصل بحل جزء من المشكلة، أما الترجمة الآلية الهجينة فزادت من تحسين المخرجات لأنظمة تمييز الكلام الآليّة (automatic speech recognition (ASR)) وأنظمة التحكم بالوسائط بالجمع بين أهم ميزات النظامين.

تتميز أنظمة الترجمة الآلية الإحصائية بقدرتها على تعلّم ترجمة العبارات وليس الكلمات المفردة فقط، وميزة أخرى لبعض هذه الأنظمة أنها تجمع عدة موارد معرفيّة لاستخدامها لتحليل الجملة المترجمة. أما نموذج الترجمة الآلية المعتمد على القواعد فيوظّف نظاما يتوفر فيه معجم غني بالشروح يحتوي بدوره على معلومات دلالية ووظيفيّة، ويُستخدم النظام لتغذية عمل الترجمة الآلية الهجينة.

يتم وصف الترجمة الآلية الإحصائية عادة بأنها اختيار الجملة الهدف التي لها أعلى احتمالية مع إعطائها الجملة المصدرية. ونماذج الترجمة المقدّمة في نظامنا هي مزيج من معاجم يتم تعلّمها إحصائيا ومُدخلة في معجم ثنائي اللغة يُستخدم في الأنظمة المعتمدة على القواعد.

النماذج اللغوية

نقدّم في هذه الورقة استخدام خصائص قواعديّة ومعجميّة في إطار عمل إحصائي، فتوظيف معلومات غنية من الناحية الإحصائية والهيكليّة في أنظمة الترجمة الآلية الإحصائية يزيد من استخدام التحليل اللغوي فيها. ويتم تدريب الأنظمة الهجينة على ترجمة أصول نصية كبيرة في لغة غير عاميّة إلى اللهجة العاميّة المطلوبة، ثم يتم استخدام الأصول النصية المولّدة لتدريب نماذج الترجمة الإحصائية، والكلمات التي ليس لها ترجمة يتم نسخها بحروف اللغة الهدف.

نماذج وظيفية

إن استخدام محددات وظيفيّة للمعلومات المعجميّة في اللغة الهدف واللغة المصدر يمنح قيمة تحليليّة إحصائيّة ودلالية أعمق للترجمة، وهذه المحددات متعددة، وبعضها يعتمد على اللغة، وقد تكون هذه الوظائف صالحة لعدة لغات أو للغة محددة فقط.

تجري ترجمة اللغة العربية الفصحى بالاعتماد على النظام المذكور باستخدام خصائص معجمية ووظيفية ونحوية، ويتم تدريب النماذج الإحصائية على أصول نصية من جمل ثنائية اللغة، أما النماذج اللغوية فيتم تدريبها على أصول نصية بلغة واحدة. وتُترجم اللغة العربية العامية باستخدام نظام الترجمة الآلية الهجين الذي يترجم اللغة العامية إلى فصحى، وتُستخدم أصول نصية ثنائية اللغة تتكوّن من جمل باللهجة العراقية والعربية الفصحى، ويمكن زيادة جودة الترجمة باستخدام خصائص متعلقة باللهجة.

إن أنظمة التحكم بالوسائط هي حلول برمجيّة تقدّم نصوصا بعدة لغات من عدة محطات

تلفزيون وراديو محلية وعالمية ونصوص محادثات هاتفية، وتدعم هذه الأنظمة الفيديوهات والصوتيات من حيث المعالجة النصية بقدرات لغوية متقدمة مثل الترجمة الآلية واستعادة المعلومات مع ترجمة الاستعلامات وغيرها، وتسهل هذه الأنظمة توليد نسخ من البث التلفزيوني والهاتفي وترجمة النص من الإنجليزية وإليها، ولكن فيما يتعلق بتمييز اللهجات العربية تكمن المشكلة الأساسيّة في صعوبة توقع نموذج لغوي إحصائي باستخدام ذخيرة لغوية (Corpus) صغيرة مقارنة بحجم المفردات، بالإضافة إلى وجود فروقات دقيقة في المفردات بين المناطق التي تستخدم في كل مكان بمعنى مختلف.

أثبتت التجربة أن النظام الهجين يكون أداؤه أفضل من النظام الإحصائي وحده أو الطرق المعتمدة على القاعدة وحدها.

Year of publication:

Hybrid Word-Character Neural Machine Translation for Modern Standard Arabic

Pamela Toman and Sigtryggur Kjartansson

الترجمة الآلية العصبية الهجينة بالكلمات والأحرف للغة العربية المعاصرة

تستخدم أداوت الترجمة الآلية العصبية التقليدية في تركيبها نهجًا يُبنى على مستوى الكلمات بافتراض أنّ كُلّ الكلمات الأساسية لعملية الترجمة لها أشكال سطحيّة متكررة ومتعددة نسبياً. هذا الافتراض ليس الأمثل لعدد كبير من اللغات غير التحليليّة التي تُبنى على تشكيل كلمات جديدة اعتمادا على العمليات المورفولوجية المعقّدة، ومن هذه اللغات اللغة العربية. لذلك قامت فكرة هذا البحث على دمج ثلاثة أنظمة بهدف توليد نظام ترجمة هجين على مستوى الكلمة والحرف، حيث اعتمد النموذج الهجين في الأساس على الترجمة على مستوى الكلمة، وتضاف إليها نماذج على مستوى الحرف تظهر فيها رموز **UNK** ، حيث يتضمن النموذج الثاني برنامج ترميز (**encoder**) لدمج الكلمات لـ**UNK** ، وتقوم وحدة فك ترميز (**decoder**) مستوى اللغة للغة الهدف بإنشاء نص للكلمات **UNK** التي تم التنبؤ بها بواسطة نموذج على مستوى الكلمة، وبهذه الطريقة المقترحة تساعد على تحقيق التوازن بين كفاءة الترجمة للمفردات المحدودة على مستوى الكلمة (**word-level**) مع المعرفة الواسعة للمصطلحات اعتمادا على تجميع وحدات الكلمة الفرعيّة (**sub-word units**) التي تم اكتسابها في مرحلة الترجمة على مستوي الحرف (**character level**). وفي مرحلة تدريب النظام المقترح، قام الباحثون بتدريب نموذج مستوى الكلمة على أزواج الجملة وبتدريب كلّ من نماذج الأحرف باستخدام عينتين، الأولى تشمل كلمات غير معروفة، والثانية عن طريق أخذ عينات الكلمات المتكررة من أزواج الجملة بمعدل مضبوط، وبهذه الطريقة لم تظهر هناك حاجة إلى خطوة منفصلة لاستبدال **UNK** كما هو الحال

في العديد من النماذج المشابهة الحالية. وطُبّق مفكك تشفير بحث في المخطط البياني للحفاظ على جمل جزئية محتملة أثناء الاستدلال. ويمكن تلخيص خطوات الطريقة المتبعة على النحو الآتي:

استخدام نموذج مستوى الكلمة إطارا لعمل فك ترميز تشفير يحسب تمثيلًا مشفرًا لكل جملة من الترميز، تقوم وحدة فك الترميز بإنشاء ترجمة من خلال التنبؤ بكلمة واحدة في كل مرة.

لتجنب تجاهل المعلومات عندما لا يكون هناك كلمات تطابق النموذج في المفردات الأساسية، قام الباحثان باستخدام أداة ترميز على مستوى الحروف متعددة الطبقات، تصبح حالتها الأخيرة المخفية تضمينًا لاستبدال التضمين بـ UNK.

يتضمن النموذج نموذجًا متكررًا منفصلاً يفك تشفيره على مستوى الحرف المعطى على مستوى الكلمة عندما يتنبأ نموذج مستوى الكلمة بـ UNK.

أثناء الاستدلال، قام الباحثان بالبحث بواسطة جميع الترجمات المحتملة باستخدام أداة فكّ تشفير البحث في الرسم البياني. وقاما بذلك عن طريق تتبع أفضل الفرضيات في كل خطوة زمنية.

ولتقييم النموذج المقترح تُرجمت النصوص المكتوبة باللغة العربية الفصحى إلى اللغة الإنجليزية، وأخذت نصوص الدراسة من مقالات إخبارية. وقد خلصت الدراسة إلى أنه بتطبيق نموذج الترجمة العصبية كان مشجّعا من الجانب النظري بسبب تحرره من التبعيّة على افتراض أن الكلمات متساوية في الطول والتكرار، وتؤدي نفس الأداء مع وقت تدريب أقل من نموذج الكلمة فقط وتحقيق أفضل النتائج للترجمة من العربية إلى الإنجليزية.

KCESS '11 Proceedings of the Second Kuwait Conference on e-Services and e-Systems, Kuwait City, Kuwait — April 05 - 07, 2011

Year of publication:2011

A proposed semantic machine translation system for translating Arabic text to Arabic sign language

Ameera M. Almasoud, Hend S. Al-Khalifa

نظام مقترح للترجمة الآلية الدلالية لترجمة النص العربي إلى لغة الإشارة العربية

عملت الأبحاث السابقة في هذا المجال على ترجمة اللغة العربية كلمة بكلمة إلى لغة الإشارة العربية متجاهلة دلالات الجمل المترجمة، ونهدف في هذا البحث إلى تحسين الأبحاث السابقة بإضافة طبقة من الدلالات بالاستعانة بتقنيات الويب (الأنطولوجيا) الدلالية، وسيقتصر نظامنا على العمل على فقه الصلاة.

تتبع ترجمة لغة الإشارة الآلية طريقتين: الطريقة المعتمدة على القاعدة والطريقة المستمدة من البيانات، وتُعرف الأخيرة أيضا بالطريقة المعتمدة على الذخيرة اللغوية (corpus)، ويمكن تقسيمها إلى ترجمة آلية إحصائية ومنهجيات الترجمة الآلية المعتمدة على المثال.(Example based machine translation(EBMT))

تتطلب الطريقة المستمدة من البيانات أصول بيانات مطلوبة مسبقا للعمل عليها، وتعتمد دقة الترجمة وجودتها على حجم أصول البيانات، أما الطريقة المعتمدة على القاعدة فتعتمد على قواعد لغوية، ولها مساران: المسار المباشر وغير المباشر، والمباشر يستخدم في القواميس ثنائية اللغة التي تتطلب ترجمة كل كلمة دون أي تحليل تفصيلي للتركيب النحوي للنص المدخل، أو أيّ علاقة بمعاني الكلمات أو العلاقات بينها، أما غير المباشر فإنه يستخدم لتحليل البناء النحوي للنص المدخل وصنع تمثيل متوسط أو مجرد له، ومن ثم توليد النص الهدف منه، وذلك يعني أن علينا تحديد بناء الكلمة وبناء الجملة والبناء الدلالي في عمليات متتالية، ونحن نستخدم

في نظامنا الطريقة المعتمدة على القاعدة.

سيحسن النظام المقترح التقنيات السابقة بالأخذ بعين الاعتبار قواعد الترجمة إلى لغة الإشارة العربية واستخدام أنطولوجيا لإنتاج ترميز لكتابة الإشارات (Sign Writing). ويتألف النظام من مجموعة من العمليات، وهي: التحليل الصرفي، والتحويل القواعدي، والترجمة الدلالية.

في عملية التحليل الصرفي نأخذ النص العربي ونرسل كل جملة للتحليل الصرفي ولعملية فك الغموض (Disambiguation) ووضع علامات أقسام الكلام (POS)، وتكون المخرجات عبارة عن سطر من الخصائص (features) يتألف من أزواج من الخاصيّة والقيمة. تأخذ عملية التحويل القواعدي النتائج السابقة كمدخلات وتطبّق عليها قواعد لغة الإشارة العربية على كل كلمة بناء على خصائصها. وتأخذ عملية الترجمة الدلالية نتائج العملية السابقة وتبحث عن كل كلمة في الأنطولوجيا المستخدمة (وهي فقه الصلاة) للحصول على رمز إشارة الكلمة، فإذا لم يكن للكلمة إشارة مقابلة، فإنّه تُستبدل بأحد مرادفاتها التي لها إشارة في قاعدة بيانات كتابة الإشارات، ثم يُستبدل كل رمز إشارة برمز الإشارة المقابل له المخزّن في قاعدة بيانات كتابة الإشارات، فإذا لم يكن للكلمة رمز مقابل في الأنطولوجيا فإنها تُهجئ بالأصابع. وسيتم تقييم النظام بناء على طريقتي تقييم، هما ترجمة مجموعة جمل آليّا ثم تدقيقها من خبراء لغة الإشارة العربية، والطريقة الثانية للطلب من الخبراء ترجمة مجموعة من الجمل يدويّا ثم مقارنة النتائج بنتائج النظام.

٣-٤-٧ أبحاث السؤال والجواب

وتضم ستة أبحاث، أربع منها نوع (أ) هي: التحديات التي تواجه اللغة العربية في المحادثة الآلية المعتمدة على النصوص مقارنة مع الإنجليزية، و نظام « QARAB» للإجابة على الأسئلة لدعم اللغة العربيّة، ورُوبوت الحوار «آليس ALICE»: التّجارب والمخرجات، ونظام السؤال والجواب AQuASys للغة العربية. وبحثان نوع (ب) هما: تصحيح استعلام بحث المستخدم العربي وتوسيعاته، وتحسين السؤال والجواب باستعمال «ووردنيت» العربيّة.

International Journal of Computer Science & Information Technology (IJCSIT) Vol 7, No 3, June 2015

Year of publication: 2015

Arabic Language Challenges In Text Based Conversational Agents Compared To The English Language

Mohammad Hijjawi, Yousef Elsheikh

التحديات التي تواجه اللغة العربيّة في المحادثة الآلية المعتمدة على النصوص مقارنة مع الإنجليزيّة

قدّمت هذه الورقة البحثية دراسة تحليليّة للمشاكل التي تواجه معالجة نصوص اللغة العربيّة باعتبارها واحدة من أشهر اللغات الطبيعيّة، ومن أكثرها استخداماً على مستوى العالم، إضافة الى توضيح أبرز المحددات الواجب مراعاتها للتعامل مع هذه اللغة انطلاقا من طبيعتها، وصعوبة بنائها، ووجود أشكال متنوعة لحروفها، علاوة على أن طريقة استخدام الكلمة بدون تشكيل تُعدُّ من المحددات للمعالجة الطبيعية للغة العربيّة.

المقدمة

نظام المحادثة (agent conversational) هو عبارة عن برنامج حاسوبي ذكي يستخدم للتعامل مع المحادثات بين المستخدم والجهاز، وقد صُمّم البرنامج واستُخدم لقياس قدرة الإنسان على تمييز ما إذا كان الطرف الآخر شخصاً أو آلة.

إنّ المراجع اللغويّة النصيّة لها ثلاثة مناهج رئيسة لبناء برامج المحادثة الإلكترونية، هي: المنهج القائم على معالجة اللغات الطبيعية، والمنهج القائم على تشابه الجمل (Sentence Similarity Measures)، ومنهج مطابقة الأنماط (Pattern Matching). ويمكن استخدام هذه المناهج للغة العربيّة واللغة الإنجليزيّة.

أنظمة المحادثة النصيّة اعتماداً على اللغات الطبيعيّة

إن ظهور برمجة اللغات الطبيعيّة أدّى إلى المزيد من التركيز على حوسبة البرامج الذكيّة لمحاكاة السلوك البشري ومعرفة الإنسان من أجل التوصل إلى استنتاجات جديدة، فـفهم الجملة يحتاج إلى فهم بنية الجملة ومكوّناتها والعلاقات بين تلك المكونات. وللغتين العربيّة والإنجليزيّة ملامح عدة تركز على مسألتين أساسيتين:

أولاً: كيفية تشكيل الكلمات والصرف الاشتقاقي (derivational morphology).

وثانياً: كيفية تفاعل الكلمات مع بعضها في بناء الجملة (inflectional Morphology).

غير أنّ تشكيل الكلمات يؤدي إلى إنشاء العديد من الجمل المختلفة، وهذا الثراء في توليد الكلمات ذات المعاني المختلفة يمثل تحدياً في اللغة العربيّة.

الأحرف الزائدة ونظام قاعدة الجذر:

إن اللغة العربيّة تتميز باعتمادها الأساسي على أصل الكلمة أو الجذر لإنشاء كلمات اللغة، لذلك، فإنّ الجذر هو النموذج الأولي الخام للكلمة التي لا يمكن إجراء مزيد من التجذيع (stemming) عليه. وهكذا فإن اللغة العربيّة غنيّة بالعديد من بنى الجذور القوية التي توفر عدداً كبيراً من الكلمات المشتقة منها. على سبيل المثال، التعرف على أنظمة أنماط الكلمات قد يؤدي إلى عدم القدرة على فهم برامج المحادثة الإلكترونية استنادا إلى برمجة اللغة الطبيعيّة، بالتالي ينعكس سلباً على أداء مهمتها.

أحرف العلة:

في اللغة العربيّة نوعان من أحرف العلة: أحرف العلة العاديّة (الألف والواو والياء) والحركات القصيرة، وتعرف بعلامات التشكيل. وعند حذف علامات التشكيل من الكلمات، قد يخلط القارئ في المعنى المراد منها. وعند بناء برامج المحادثة الإلكترونية باستخدام اللغة العربيّة الحديثة، سيؤدي التشكيل الى زيادة في التعقيد، بخلاف اللغة الإنجليزيّة التي لا تتضمن علامات تشكيل.

الأسماء والأفعال:

إن للأسماء في اللغة العربيّة ثلاث حالات: المفرد والمثنى والجمع. وعند التفريق بين المذكر والمؤنث تُضاف الضمائر، على عكس اللغة الإنجليزيّة التي لاتحتاج لذلك. إنّ هذه الضمائر لا تعكس إذا كان المقصود إنسانا أو جمادا، وهذا يزيد من التحدي القائم على الكشف والتحليل والحكم على معاني كلمات اللغة العربيّة. بينما تستخدم الأفعال في اللغة العربيّة للتعبير عن المذكر والمؤنث، وهذا ماتفتقده اللغة الإنجليزيّة. ويمكن أيضاً استخدام الأفعال للتعبير عن عدد من الأشخاص المعنيين للقيام بعمل ما، مما يولد صعوبة في تحليل هذه الأفعال وتحديد ما إذا كان المقصود فيها مفرداً أو مثنى أوجمعاً.

الأسماء المعرّفة والكلمات الأجنبية (الدخيلة):

تشمل الأسماء المعرّفة باللغة العربيّة اسم شخص معيّن أو حدث أو تاريخ، والمقصود بالكلمات الأجنبيّة الكلمات التي أقتبُست من اللغات الأخرى. تتمثل صعوبة هذه الكلمات بعدم امتلاكها لأيّ جذر ينتمي للغة العربيّة، ولذلك ينبغي على أنظمة المحادثة النصية التعامل مع هذه الكلمات وتقبُّلها كما هي.

نستنتج مما سبق أن برمجة اللغات الطبيعيّة لاتزال تعاني من العديد من القيود والتحديات. فعلى سبيل المثال، كلمات اللغة العربيّة قد يكون لها أكثر من معنى، أو قد تكون هناك كلمة واحدة تستخدم للتعبير عن أكثر من شيء. والمحادثات الحقيقيّة بين الأشخاص في كثير من الأحيان تكون مليئة بالأخطاء النحْويّة مما يزيد من الوقت اللازم لمعالجتها.

وتتميز اللغات الغنية مثل اللغة العربيّة بالاشتقاقات ، بازدياد ملحوظ في صعوبة بناء أنظمة المحادثة النصية لها. ومن المفترض أن يكون النظام قادراً على تفسير الكلام واتخاذ الإجراءات اللازمة للرد على هذا الكلام.

أنظمة المحادثة النصية اعتماداً على مطابقة الأنماط:

إن النص المعتمد على مطابقة النمط في علوم الحاسوب هو عملية البحث عن تسلسل في جزء من النص للبحث عن كافة حالات وجود هذا التسلسل داخل نصّ معيّن. وهذه الآلية تستخدم خوارزمية معيّنة لمعالجة محادثات المستخدمين من خلال مطابقة أنماط أنظمة المحادثات النصية مع كلام المستخدم. يُعدُّ نهج مطابقة الأنماط واحداً من أنجح الطرق لتطوير أنظمة المحادثات النصية، ولهذه التقنية العديد من المزايا، بما في ذلك سهولة فهمها لأنّها لغة مستقلة و مناسبة لكلا اللغتين الإنجليزية والعربية. ولا تتطلب هذه الآلية مراحل معقدة في تحليل الكلام، وبالتالي فإنّها ليست مكلِّفة حاسوبياً. إنّ الأنظمة المستخدمة بهذا المنهج قادرة على دعم المحادثات بشكل فعال ولأعداد كبيرة من المستخدمين. كما تستطيع هذه الأنظمة حلّ العديد من التحديات اللغويّة التي تواجه تقنية برمجة اللغة الطبيعيّة. ولعلّ العيب الرئيسي لهذا المنهج هو حاجته لعدد كبير من الأنماط لبناء مجال متماسك، وتأتي الحاجة لهذا العدد الكبير من الأنماط من العديد من القضايا؛ أولاً: وجود العديد من الطرق المتنوعة التي يمكن للمستخدم توظيفها لبناء الجمل.

ثانياً: الطبيعة القوية والغنية للغة تجبرها على تغطية جميع التغييرات المتوقعة للكلمات لتلبية حالات مختلفة مثل المفرد/ الجمع. وهذا العيب يواجه كلا اللغتين (الإنجليزيّة والعربيّة)، لكنّه أعظم بالنسبة للغة العربيّة.

أنظمة المحادثة النصية اعتماداً على مقاييس تشابه الجمل:

تقوم تقنيات تشابه الجمل على قياس مستوى التشابه بين الجمل، ويجري التركيز على هذا المنهج باعتباره منهجا رئيسيّا لبناء أنظمة المحادثة النصية. ولقياس التشابه بين الجمل يُستخدم منهجان رئيسان هما " التحليل الدلالي الكامن ((Latent Semantic Analysis" (LSA)) وتشابه الجملة استناداً إلى الشبكات الدلالية (semantic networks)، وإحصائيات الذخيرة اللغويّة.

يقوم المنهج الأوّل بقياس التشابه بين الكلمات من خلال استخدام الحسابات الإحصائية،

ومن ثم، واعتماداً على هذه الحسابات، يولّد تمثيلا لتشابه الكلمات والمقاطع في النص، وذلك من خلال إنشاء مصفوفة من الكلمات استناداً إلى عدد التكرارات التي ظهرت فيها كلمة معيّنة في سياق محدد دون النظر لترتيب الكلمات في الجملة. ولا يتعامل هذا المنهج مع العلاقات النحْويّة للكلمات، وبالتالي، فإنّ هذا يسبب مشاكل تؤدي إلى عدم القدرة على تحليل الجمل بشكل صحيح، إضافة إلى كونها تقنية مكلفة حاسوبياً. أما المنهج الثاني، فهو تقنية تقوم على حساب التشابه بين الجمل من خلال استخدام قاعدة بيانات معجميّة وضعت في جامعة برينستون. وتصنف قاعدة البيانات الكلمات الى أربع فئات رئيسة (الأسماء والأفعال والصفات والحالات). وقد بدأ العمل البحثي لهذا المنهج المختص باللغة العربيّة في عام ٢٠١٢. ونتيجة لذلك، فإنّ هذا المنهج لايزال جديداً ويحتاج إلى مزيد من الأبحاث للاعتماد عليها بشكل رئيسي لبناء البرمجيات اللازمة.

الخلاصة

في هذه الورقة البحثيّة، جرى تقديم وصف موجز للغة العربيّة مقارنة باللغة الإنجليزيّة. وركّزت هذه الورقة على تعقيدات اللغة العربيّة وتحدياتها في العمل القائم على الحوسبة بشكل عام. إضافة إلى ذلك، فإنّ هذا العمل البحثي ركز بشكل خاص على مناقشة المناهج المستخدمة لبناء أنظمة المحادثة النصية. هناك ثلاثة مناهج رئيسة مستخدمة لبناء أنظمة المحادثة النصية وهي: مطابقة الأنماط، ومعالجة اللغة الطبيعيّة، ومقاييس تشابه الجملة. وقد بيّنت هذه الورقة البحثيّة المزايا والعيوب لكل منهج استُخدم في بناء هذه الأنظمة، فقد ناقشت هذه الورقة بإيجاز المناهج الثلاثة والتحديات والقيود في بناء الأنظمة للغة العربيّة مقارنة باللغة الإنجليزيّة. وقد ظهر أن اللغة العربيّة أكثر تحدياً من اللغة الإنجليزيّة لأسباب كثيرة. أولها، تعقيد اللغة العربيّة بوصفها لغة ساميّة غنية بالمشتقات وغيرها من التعقيدات المستمدة من جوانب مختلفة. وثانيها، أن معظم البحوث في مجالات برمجة اللغات الطبيعيّة منصبة على اللغة الإنجليزيّة. وقد أدى ذلك الى ازدياد التحدي والحاجة الى مزيد من البحوث في هذه المجالات للغة العربيّة قبل البدء والشروع ببناء أنظمة المحادثة النصية.

SEMITIC '02 Proceedings of the ACL-02 workshop on Computational approaches to semitic languages, pages 1- 11

Year of publication: 2002

QARAB: A Question Answering System to Support the Arabic Language

Bassam Hammo, Hani Abu-Salem, Steven Lytinen

نظام " QARAB" للإجابة على الأسئلة لدعم اللغة العَربيّة

المقدمة

تصف هذه الدراسة تصميم نظام للرد الآلي على الأسئلة وتنفيذه، أطلق عليه اسم (QARAB). وهو نظام يأخذ أسئلة اللغة الطبيعيّة المعبّر عنها باللغة العربيّة، ويحاول تقديم إجابات قصيرة عليها. والمورد الأساسي لمعلومات هذا النظام عبارة عن مجموعة أو قائمة من النُصوص الصّحفية العربيّة المستخرجة من مجلّة الرّاية القطرية.

خلال السنوات القليلة الماضية، استطاع قطاع استرجاع المعلومات (Information Retrieval) التعامل مع هذه المشكلة من جهة اللغة الإنجليزيّة باستخدام تقنيات استرجاع المعلومات القياسيّة، إلا أنه حقق نجاحاً مُتواضعاً. في هذه الورقة، نعالج هذه المشكلة من جهة اللغة العربيّة باستخدام تقنيات استرجاع المعلومات التقليديّة، إلى جانب تقنية معالجة اللغة الطبيعيّة المُتطوّرة. ولتحديد الجواب، اعتمدنا استراتيجية مطابقة الكلمات الرّئيسيّة (matching simple structures) جنباً إلى جنب مع مطابقة الهياكل البسيطة المستخرجة من كلّ من السؤال والوثائق التي يختارها نظام استرجاع المعلومات المستخدم. ولتحقيق هذا الهدف استخدمنا نظام الوسم (Tagger) الموجود لتحديد الأسماء المناسبة وغيرها من العناصر المعجمية المهمة، وبناء الإدخالات المعجمية لها بشكل آليّ (تلقائي). وقُمنا أيضاً بإجراء تحليل لأشكال الأسئلة العربيّة ومحاولة فهم أفضل أنواع الإجابات التي يرتضيها المستخدمون.

ويمكن تلخيص الأسلوب الذي تقترحه هذه الدراسة على النحو التالي: يعالج نظام استرجاع المعلومات السؤال المطروح من خلال عملية الاستعلام في محاولة لتحديد الوثائق المرشّحة التي قد تتضمن الجواب؛ ثم تُستخدم تقنيات معالجة اللغة الطبيعيّة لتحليل السؤال وتحليل أعلى الوثائق رُتبة، وتمّ استرجاعها بواسطة نظام استرجاع المعلومات.

مشكلة البحث

لا تزال معالجة اللغة الطبيعيّة فيما يخص اللغة العربيّة في مراحلها الأوليّة، مقارنة بالعمل على اللغة الإنجليزيّة، التي استفادت بالفعل من البحوث المكثّفة في هذا المجال. غير أنّ هناك بعض الجوانب التي تُبطّئ مسار التقدم في مجال معالجة اللغة العربيّة الطبيعيّة مقارنة باللغة الإنجليزيّة واللغات الأوروبيّة الأخرى. فبالإضافة إلى ما ذُكر أعلاه من مسائلَ لغويّة، ما زال هناك نقص في متون اللغة العربيّة، والمعاجم والقواميس المقروءة آليّاً، التي تُعدّ ضرورية لدفع البحوث في مجالاتٍ مُختلفة.

لقد شهدت السّنوات الأخيرة زيادة ملحوظة في حجم البيانات المتاحة على الشّابكة. وغالباً ما يكون لدى المستخدمين أسئلة مُحدّدة في أذهانِهم، ويأملون في الحُصول على إجابات عليها. كما أنهم يرغبون في أن تكون تلك الإجابات قصيرة ودقيقة، كما يُفضّلون دائماً التعبير عن الأسئلة بلُغتهم الأصلية دون الاضطرار للّجوء للغة استعلام مُعيّنة أو لقواعد محددة، أو حتى استخدام نطاق أو مجال مَعرِفة مُحدّد. أما الأسلوب الجديد الذي اعتُمد هنا، ليوائم احتياجات المستخدِمين، فيأتي لغرض إجراء تحليل فِعليّ للسؤال من وُجهة نظر لُغوية ومُحاولة فهم ما يعنيه المستخدم فعليّاً. إنّ نظام (QARAB) هو حصيلة اقتران تقنيات استرجاع المعلومات التقليديّة (IR) بتقنية مُعالجة اللغة الطبيعيّة المتطورة (NLP).

نظام QARAB

إن الهدف الرئيسيّ من نظام QARAB هو تحديد المقاطع أو الفقرات النّصيّة التي تُجيب على

سؤال مكتوب باللغة الطّبيعيّة. ويُمكن تلخيص مهمة النّظام كالآتي: إذا كان لدينا مجموعة من الأسئلة التي جرى التعبير عنها باللغة العربيّة، نقوم بإيجاد إجابات على تلك الأسئلة في إطار الفرضيّات التالية:

الجواب موجود في مجموعة من النّصوص الصّحفية العربيّة المستخرجة من صحيفة الرّاية القطرية.

عدم تشتت الإجابة بين عدة وثائق (أي أن تكون جميع المعلومات الداعمة للإجابة موجودة في وثيقة واحدة).

أن تكون الإجابة على شكل فقرة قصيرة.

وتجري المعالجة الأساسيّة لعمليّة الإجابة على السؤال عبر ثلاث خطوات رئيسية:

معالجة السؤال المدخل

استرجاع الوثائق المرشّحة (الفقرات) التي تحتوي على إجابات من نظام استرجاع المعلومات.

معالجة كلّ واحدة من الوثائق المرشحة (الفقرات) بالطريقة نفسها التي عُولج بها السؤال عنها واستعادة الجمل التي قد تحتوي على الجواب.

وسيجري تقييم نظام «QARAB" من خلال مجموعة واسعة من الأسئلة التي يطرحها مستخدمو اللغة العربيّة أثناء مرحلة الاختبار والمراحل النهائية. وسيقوم المستخدمون أنفسهم بعد ذلك بتقييم ما إذا كانت الإجابات التي يصدرها النظام مُقنعة أو لا.

الشكل الأساسي للمُعالجة ضمن نظام استرجاع المعلومات (IR)

أولاً: مُعالجة الوثيقة أو المُستَند

تُعدُّ هذه الخطوة ضرورية لنظامنا، أولاً: لأنّ المقالات الصّحفية المأخوذة من صحيفة الراية القطرية محفوظة بصيغة نصّية باستخدام برنامج (Arabic Windows ١٢٥٦ encoding

scheme). وهو مُصمّم لاستخراج جميع علامات الوسم بلغة ترميز النص التشعّبي (html)، والحصول على محتويات النص النقيّة للمقالة؛ ثانياً: لقد أُنشئ نظام استرجاع المعلومات (IR) باستخدام نموذج قاعدة البيانات العلائقيّة. وتتضمن هذه الخطوة التحليل اللغوي، وإزالة الكلمات الشاذة، واستخراج جذور الكلمات، وتوزين المفردات.

ثانياً: استخراج جذر الكلمة

لاستخراج جذور الكلمات العربيّة من كلماتها يقوم محلل الجذور بإجراء العمليات التالية على كل كلمة وفقاً للترتيب التالي:

إزالة «الـ التعريف».

إزالة حروف العطف.

إزالة «اللواحق".

إزالة «السوابق».

مطابقة النمط.

معالجة السُؤال في نظام QARAB

يتطلّب الوصول إلى فهم السؤال استخدام مُعالجة دلاليّة عميقة، وهي ليست مهمة سهلة في معالجة اللغة الطبيعيّة. وفي الواقع، لا يوجد في معالجة اللغة العربيّة الطبيعيّة أبحاث يُعتدّ بها على المستوى الدلالي. لذلك، يستخدم نظام QARAB مستوىً سطحيًا لفهم اللغة لمعالجة الأسئلة، كما أنّه لا يُحاول فهم محتوى السُؤال بمستوى دلالي عميق. ويعالج نظام QARAB السؤال الوارد كـ "حقيبة من الكلمات» (bag of words) ويُبحث في ملف فهرسها للحصول على قائمة من الوثائق المرتبة التي قد تحتوي على الجواب.

مُعالجة الجواب في نظام QARAB

إنّ المدخلات إلى أداة توليد الجواب أو تكوينه في نظام QARAB هي السؤال المطروح

باللغة الطبيعيّة ومجموعة صغيرة من الوثائق المرتبة لهذا الغرض. يُعالج السؤال أولاً عن طريق وضع علامات (وسم) على جميع الكلمات، ومن ثم تُسترجع مجموعة الوثائق ذات الصلة التي قد تحتوي على الجواب من قبل نظام استرجاع المعلومات. في عملية إنتاج الجواب، تُجمع كل المقاطع أو الفقرات في الوثائق أو المستندات ذات الصلة، التي تُطابق تماماً "حقيبة الكلمات" لأجل استخدامها في مزيد من عمليات المعالجة اللاّحقة. وتشمل مناطق الإجابة عادة معظم المصطلحات التي تظهر في الاستعلام الأصلي، بالإضافة إلى الأسماء الصّحيحة التي يجب أن تظهر في الإجابة النهائية.

الخاتمة

تعرض هذه الدراسة وصفاً لتصميم أسلوب لنظام رد على الأسئلة ولتنفيذه. ويقدّم هذا الأسلوب إجابات مختصرة على الأسئلة المكتوبة أو المعبّر عنها باللغة العربيّة. يستخدم هذا النظام كُلا من تقنيات استرجاع المعلومات ومعالجة اللغة الطبيعيّة لمعالجة مجموعة من الوثائق النصية العربيّة، باعتبارها المورد الرئيسيّ للمعرفة أو للمعلومات. وقد طُبّق النظام الذي يُسمّى "QARAB" بشكل فعلي، ويبدو أن التّحليل الأوليّ يُشير إلى نتائج واعدة. ويقتصر نجاح النظام في إطاره العام على حجم الأدوات المتاحة التي تُطوَّر للتعامل مع اللغة العربيّة. وما زال العمل جاريا لإدماج عمليّة استرجاع المعلومات في النظام وتوسيع نطاق وظائف نظام معالجة اللغة الطبيعيّة من خلال إيجاد خوارزميّات أكثر تطوراً لإنتاج إجابة موجزة في الوقت المناسب.

Comp. y Sist. 2015, vol.19, n.4, pp.625-632. ISSN 1405-5546. http://dx.doi.org/10.13053/CyS-19-4-2326.

Year of publication: 2015

ALICE Chatbot: Trials and Outputs

Bayan AbuShawar, Eric Atwell

رُوبوت الحوار "آليس ALICE": التّجارُب والمُخرَجات

المُقدمة

روبوت الحوار (chatbot) هو أداة مُحادثة تتفاعل مع المستخدمين باستخدام اللغة الطبيعيّة، وهناك العديد من روبوتات الحوار المتاحة للعمل في المجالات المختلفة، غير أنّ تشفير قاعدة معرفة أو معلومات روبوتات الحوار يتم بطريقة يدويّة (coded hand) في أدمغتها. وتُقدّم هذه الورقة البحثيّة لمحة عامة حول روبوت الحوار «آليس ALICE"، والصيغة التي يعمل من خلالها وهي "AIML"، والتجارب التي أجريناها عليه من أجل إنتاج نماذج مُختلفة من "آليس" تعمل بشكل تلقائي استناداً إلى أسلوب الذخيرة اللغوية. وأوردنا في هذه الدراسة وصفاً للبرمجيات التي طُوِّرت، وتعمل على تحويل النّص (الذخيرة اللغوية) القابل للقراءة إلى صيغة «AIML"، بالإضافة لوصف الذخائر اللغويّة المختلفة التي استخدمناها. وأظهرت تجاربنا إمكانية إيجاد نماذج مفيدة دون الحاجة إلى استخدام تقنية معالجة اللغة الطبيعيّة المتطورة أو تقنيات التعلم الآلي المعقّدة. وقد استخدمت هذه النماذج الأولية باعتبارها أدوات لتمرين لغات مختلفة لوضع تصور أو رؤية للمتن وتقديم إجابات للأسئلة.

مشكلة البحث والمُبررات

لقد جرى التعامل مع روبوت الحوار المفتوح المصدر "آليس ALICE" منذ عام ٢٠٠٢. و»آليس ALICE" هو اسم يدل على اختصار لعبارة: "كيان أو كينونة حاسبة لغويّة اصطناعيّة

على الشابكة»، أنشأه «والاس Wallace" في عام ١٩٩٥ . ومن حيث بناء «آليس» هناك فصل تام وواضح بين «مُحرّك روبوت الحوار" و «نموذج المعلومات اللغويّة» (language knowledge model) ، وبالتالي يمكن توصيل نماذج المعرفة اللغويّة البديلة وتشغيلها من خلاله.

من جهة أخرى، هناك فرق رئيسي آخر بين أسلوب «آليس» وغيره من أدوات روبوتات الحوار مثل (AskJeeves) وهو البساطة المدروسة للخوارزميات المطابقة للنمط (pattern-matching algorithms)؛ في حين يستخدم روبوت الحوار (AskJeeves) تقنية معالجة اللغة الطبيعيّة المتطوّرة بما في ذلك التّحليل الصّرفي-النّحوي، وتحليل النّص اللغوي، والتّحليل البُنيوي الدلالي (semantic structural analysis). ويعتمد آليس على عدد كبير جداً من «الفئات أو التّقسيمات» الأساسيّة أو قواعد مُطابقة أنماط المدخلات مع قوالب المخرجات. كذلك يسعى «آليس» نحو الحجم أكثر من سعيه نحو التطوّر، بحيث يُعوّض عدم وجود وحدات لمعالجة اللغة الطبيعيّة من النّاحية الصّرفية والنّحوية والدلاليّة من خلال الاستعانة بعدد كبير جداً من القواعد البسيطة، فهو يحتوي على نحو خمسين ألف فئة، كما أننا قمنا بتطوير إصدارات أكبر، تصل إلى أكثر من مليون فئة أو قاعدة.

ولدينا تقنيات لتطوير نماذج لُغويّة جَديدة لروبوت الحوار «آليس» التي تُعالج موضوعا مُعيّنا أو مُتخصّصا، حيث تنطوي هذه التّقنيات على أسلوب التعلّم الآلي من الذخيرة اللغويّة، والحوارات الناتجة عن أسلوب التّدريب الأساسي.

وتضمنت هذه الدراسة المباحث الرئيسية التالية:

بناء نظام «آليس»

يُخزّن "آليس" المعلومات حول أنماط المحادثة الإنجليزية على شكل ملفات AIML. والمصطلح «AIML" هو اختصار لحروف كلمات "لُغة ترميز الذكاء الاصطناعي"Artificial Intelligence Mark-up Language)) المُشتقة من لُغة التّرميز المُوسّعة (Extensible Mark-up Language XML).

ويتكون ملف AIML من كائنات البيانات التي يُسمّى الواحد منها «كائن AIML"، وتتكون من وحدات، هي الموضوعات والفئات. أما الموضوع فهو عنصر اختياري عالي المستوى، وله سمة اسم ومجموعة من الفئات ذات الصلة بهذا الموضوع. أما الفئات فهي الوحدة الأساسية للمعلومات في AIML. وتُشكّل كلّ فئة قاعدة لمطابقة المدخلات وتحويلها إلى مُخرجات، وتتكون من نمط (يُمثّل مُدخَلات المستخدم) ونموذج يحتوي على جواب الروبوت «آليس». ويُعدُّ نمط AIML بسيطا، حيث يتكوّن فقط من الكلمات والمسافات ورموز حرف البدل "_" وإشارة النجمة»*". وقد تتكوّن الكلمات من حروف وأرقام، بدون أيّة حروف إضافيّة أخرى، ويتمّ فصل الكلمات عن طريق مسافة واحدة، وتقوم حروف البدل بوظيفة الكلمات.

وهناك ثلاثة أنواع للفئات: الفئات الذّريّة (Atomic categories) (وهي الفئات التي تشتمل على أنماط بدون حروف البدل وإشارة «_» و»*»)؛ والفئات الافتراضية (Default categories) (وهي الفئات التي تشتمل على حروف البدل وإشارة "_" و"*") والفئات التكرارية (Recursive categories) (وهي الفئات التي تشتمل على قوالب تتضمن علامات (<srai> و <sr> ، وترمز على التوالي إلى: ذكاء اصطناعي مُتكرر ببساطة، واختزال رمزي).

ومن جهة تقنية مطابقة نمط AIML/ALICE، يحاول مُترجم AIML مُطابقة الكلمة بالكلمة للحُصول على أطول مطابقة للنمط، وهذا هو الأفضل عادةً.

تكوين قاعدة معرفيّة بصيغة AIML بشكل تلقائي

قمنا بتطوير برنامج يعمل بلغة جافا يُحوّل النّص القابل للقراءة إلى صيغة نموذج لغة روبوت المحادثة، وهناك إصداران من البرنامج، يعتمد الأوّل على فئة قالب النمط البسيط، في حين يمثّل الثاني القالب الذي يحمل إجابة الروبوت. وعادة، تحتوي ذخيرة الحوار على تعليقات إيضاحيّة لغويّة تظهر أثناء المحادثة المنطوقة مثل التداخل، وتستخدم بعض الحشو اللغويّ. وللتعامل مع الشروح اللغويّة والحشو، يتكوّن البرنامج من أربع مراحل:

الأولى: قراءة نص الحوار من المتن وإدراجه ضمن نواقل.

الثانية: وحدات إعادة مُعالجة النصوص، حيث تتم تصفية جميع الشروح اللغويّة مثل التداخل(overlapping)، والحشو (fillers)، والتعليقات اللغويّة (linguistic annotations) الأخرى.

الثالثة: وحدة التحويل، حيث يتم تمرير النص المعالَج مُسبقاً إلى المحوّل بحيث تُعدُّ الجولة الأولى نمطا والثانية قالبا. ويتم خلال هذه المرحلة إزالة جميع علامات الترقيم من الأنماط وتحويلها إلى حروف كبيرة.

الرابعة: نسخ هذه الفئات الذرية في ملف AIML.

أما الإصدار الثاني من البرنامج فيعتمد على أسلوب أكثر عمومية لإيجاد أفضل صِنو أو مشابه لمدخلات المستخدم من الحوار.

تعلم الدّردشة بالاعتماد على ذخائر لغوية مُختلفة

في إطار العمل على تعزيز نظامنا وتطويره، قمنا بتجربة أنواع مختلفة من الذخائر اللغويّة: الحوار، المونولوج، الخ. وعرضت من خلال هذا المبحث مناقشة موجزة حول جميع الذخائر اللغويّة المستخدمة وحول كيفيّة تطوير البرمجيات، وكان أبرزها:

استخدام ذخائر حواريّة مختلفة بلغات مُختلفة.

استخدام ذخائر مو نولوجيّة ثنائيّة اللغة.

استخدام ذخائر سؤال وجواب/ أسئلة مُتكرّرة.

الخاتمة

روبوت الحوار هو أداة محادثة تتفاعل مع المستخدمين باستخدام اللغة الطبيعيّة، وقد استعرضت هذه الورقة روبوت الحوار «آليس» من جهة قاعدة المعلومات التي يعتمد عليها وتقنية مطابقة نمطها. لكن العيب الرئيسيّ في "آليس" وغيره من روبوتات الحوار هو أن تطوير موارد معرفتها يُجرى بشكل يدوي وليس آليا. لقد تمكّنا هنا من إنشاء برنامج يقرأ

من الذخيرة اللغويّة ويحولها إلى قاعدة معلومات لـ «آليس». وقد استخدمت ذخائر مختلفة لإعادة تشكيل «آليس» من جديد، بجيث يكشف عن تطبيقات أخرى مفيدة من روبوتات الحوار بدلاً من اقتصار استخدامها أداة للترفيه. ويمكن استخدام روبوت الحوار أداةً لتحريك الجسم أو تصويره، ولتعلّم اللغة الإنجليزيّة أو العربيّة، أو اللغات الأخرى، والوصول إلى بوابة المعلومات لتقديم إجابات على الأسئلة.

لقد تمكّنا من إثبات أن مُحرّك روبوت الحوار بسيط على غرار «آليس" يمكنه أن يعطي نتائج ذات تقدير جيد، وعلى الأقل أسوةً بتلك التي تعطيها مُحرّكات البحث الأكثر انتشارا الموجودة على الشابكة وتعمل بصفة تجارية. ولم نكن في هذه المهمة بحاجة إلى استخدام تقنية تحليل اللغة الطبيعيّة المتطوّرة أو الاستدلال المنطقي، واكتفينا بمجموعة بسيطة (لكنها كبيرة) من قواعد مُطابقة النماذج-الأنماط.

Recent Advances in Applied Computer Science and Digital Services. ISBN: 978-1-61804-179-1, 2011

Year of publication: 2011

AQuASys: A Question-Answering System For Arabic

Sman Bekhti, Maryam Al-harbi

نظام السؤال والجواب AQuASys للغة العربيّة

تمثّل منظومات الأسئلة والأجوبة حلولاً لمسألة استرجاع المعلومات والمشاركة في المعرفة واستكشافها. وقد وُضعت كثير من هذه الأنظمة للغات أخرى غير العربيّة ووّضعت بعض هذه الأنظمة للأسئلة للحصول على الأجوبة من الشابكة (الإنترنيت).

صُمّم نظام **AQuASys** ليمكّن المستخدم من تقديم سؤال باللغة العربيّة العاديّة ويحصل على جواب دقيق لذلك السؤال. وقد صُمّم هذا النظام ليخدم أيّ كيان (**entity**) مثل أسماء الأشخاص أو المواقع أو المؤسسات أو الزمن أو الكميّات وغيرها.لذلك فإنّ هذا النظام يقبل الأسئلة التي تبتدئ بأدوات الاستفهام (من وما وأين ومتى وكم للعدد وكم للكمية). وقد قادنا البحث المعمّق لتراكيب الجمل إلى أنّه يجب أخذ تراكيب الكلمات بعين الاعتبار أيضًا. وقد افترضنا أنّ الجواب هو عبارة قصيرة، ولا يمكن الحصول عليه إلا من وثيقة واحدة وليس من وثائق متعددة.

وقد جرى تقييم كفاءة النظام عن طريق الفحص من متكلمين باللغة العربيّة ممن لغتهم الأمّ هي العربيّة، وقد استعملت مقاييس خاصة لتقييم الأجوبة المتحصّلة. هذا بالإضافة إلى وجود وحدة للتحليل الصرفي (**Morphological analysis module**) التي تقوم بتحليل صيغ الأسئلة وتراكيبها باللغة العربية بالإضافة إلى تعريف وسائل تقييم عددية لتحديد دقة الأجوبة.

بنية النظام

بُني النظام من أربع وحدات (modules) هي: تحليل السؤال، وترشيح الجمل، وتقييم الأجوبة رقميًا، ووحدة الترتيب. وقد وُصف كل جزء من هذه الأجزاء الرئيسية بناء على أجزائها الفرعية.

وحدة تحليل الأسئلة

هذا الجزء مهم وذو تأثير بالغ على أداء النظام. يقوم هذا الجزء بمعالجة السؤال بهدف تحديد عنصرين مهمين، هما نوع الجواب المتوقّع، والمعلومات الأساسية التي يحتويها السؤال:

تحديد نوع الجواب المتوقع.

أجريت دراسة مستفيضة لتركيبات الجمل الاستفهاميّة في اللغة العربيّة، واستُنتجت جملة من القواعد لتمييز أنواع الأسئلة واحتماليّة أجوبتها. إنّ معظم أنظمة الأسئلة والأجوبة العربيّة عالجت فقط أسماء الاستفهام المعروفة مثل من وأين ومتى وكم العدديّة وكم الكميّة.

في اللغة العربيّة يمكن صياغة هذه الأسماء بأشكال أخرى لها المعنى نفسه. فمثلا يمكن الاستعاضة عن (أين) بالعبارة (في أي بلد أو في أي مدينة) وهكذا. والحال نفسه في أسماء الاستفهام الأخرى، وعلى هذا فإنّ السؤال نفسه يمكن توجيهه بأشكال عديدة ممكنة، ولا يمكن التغاضي عن الصيغ المختلفة والتركيز على صيغة واحدة.

المهمة الرئيسيّة الثانية هي تجزئة سؤال المستخدم، وتتكوّن هذه المهمة من ترتيب كلمات السؤال بثلاثة أصناف: اسم الاستفهام وفعل الجواب والكلمات المفتاحيّة للسؤال.

إن تمييز اسم الاستفهام يسهّل استرجاع الجواب، وهذا مهم للتعرف على الجملة التي تحوي الجواب المتوقع. والفعل إن وُجد هو إحدى الكلمات المهمة في أيّ سؤال أو جملة، لذلك فإنّ التعرف على الفعل ضمن السؤال أو الجملة هو أمر مهم، فالفعل باعتباره إحدى الكلمات المفتاحيّة يُعطى وزنًا أكبر في ترشيح الجملة ذات العلاقة. وبالطبع فإن التعرّف على الفعل في

جملة السؤال يعتمد على تركيبة السؤال وتحليل الجملة في اللغة العربيّة. أما الكلمات المفتاحيّة فهي كلّ الكلمات الواردة في سؤال المستخدم من غير الأفعال وأسماء الاستفهام. ويمكن استخلاص هذه الكلمات من السؤال من خلال مرحلة الترشيح (filtering).

زيادة الكلمات المفتاحية للسؤال:

لغرض الزيادة في الدقة، هناك حاجة لتوليد كلمات مفتاحيّة تُضاف للكلمات المفتاحيّة الواردة في السؤال، وتتم زيادة هذه الكلمات بالاعتماد على طبيعة السؤال. فمثلاً إذا كان السؤال عن موقع (مثلاً أين موقع المدينة الفلانيّة) تُضاف كلمة ذات علاقة بالمواقع مثل مكان، موقع ، بلد ..الخ

وحدة ترشيح الجمل ذات العلاقة

هذه الوحدة هي وحدة رئيسيّة تعمل على التعرّف على الجمل الرئيسيّة الأكثر علاقة مع سؤال المستخدم، فهي تستعمل كلّ المعلومات المستخلصة من وحدة تحليل السؤال، وذلك للتعرف على الجمل التي يُحتمل أن تحتوي على جواب السؤال أكثر من غيرها. هذه الجمل يجري ترتيبها بالتسلسل حسب أهميّتها وفق عامل الأهميّة.

هذا بالإضافة إلى أنّه بعد تحليل السؤال، فإنّ النظام يقوم بعملية إيجاد جذع الكلمة (technique stemming) للكلمات المفتاحية في السؤال والكلمات المفتاحيّة في الوثائق. وهذا ضروري للتغلب على التغييرات النحويّة في الكلمات، ومن ثم تحديد موقع الجمل في الوثيقة التي تستعمل طريقة مطابقة السلسلة (string matching approach). في هذه المرحلة تُعرف الجملة ذات العلاقة بالاستناد إلى وجود الكلمات المفتاحيّة في سؤال المستخدم في الجملة نفسها. هذه الكلمات المفتاحيّة تتضمن الأفعال في السؤال، سواء أكانت بشكلها المجرّد أم مع سوابقها ولواحقها في الكلمة أثناء عملية معالجة الوثائق. وكخطوة أوليّة، تقلّص الجمل ذات العلاقة نتيجة ترشيح الجمل إلى حد أنّ الجملة المختارة تحتوي على كلمة مفتاحيّة واحدة (سواء مع السوابق واللواحق أو بدونها).

تكامل جذع الكلمة في اللغة العربية

لأنّ الوثيقة تكون في الغالب خالية من الكلمة في السؤال كما هي (مع السوابق واللواحق)، فإنّ تجزئة الكلمات إلى سوابق وجذع ولواحق يصبح أمرًا لا مناص منه. فمثلا السؤال: «أين وُلد الخوارزمي؟» وُجد أن هناك وثيقة تحوي عبارة: « الخوارزمي مولود في مدينة خوارزم في خراسان». فإذا لم تستخدم عملية تجزئة الكلمة فإنّ هذه الجملة لن تُعدُّ ذات علاقة. أما إذا ما اعتبر الفعل «ولد» ومشتقاته «مولود» اعتمادًا على التحليل الصرفي الذي تمتاز به اللغة العربيّة فسيكون بالإمكان اعتبار هذه الجملة هي الجملة التي تجيب على السؤال.

لقد استخدمنا طريقة التحليل الصرفي المعروفة « محلل خوجا الصرفي»، وهو يقوم بالعمليات الآتية:

حذف ألف لام التعريف.

حذف واو العطف.

حذف السوابق.

حذف اللواحق.

مطابقة الهيئة (**pattern matching**).

كما يعطي المحلل الصرفي كذلك جذر الكلمة الذي يمكن اعتباره كلمة مفتاحيّة تقارن مع غيرها.

وحدة ترتيب الأولويات

حينما نحصل على جمل مرشّحة لأن يكون فيها جواب للسؤال، يبدأ النظام بترتيب هذه الجمل وفق خطّة ترتيب محددة، حيث تُرتّب الجمل وفق نسب دقتها. في هذه الوحدة يتقرر ما هي أفضل جملة مرشحة لأن يكون فيها الجواب. لقد طُوّرت آلية وفق دقة انطباق الجملة مع السؤال باستعمال قوانين مستندة على ألية قياس مقدار التشابه. وقد بيّنت كثير من الدراسات أنّ

تطبيقات السؤال والجواب تحتاج إلى التعرّف على التشابه بين السؤال مع الجواب، وبين السؤال مع السؤال. تستند هذه العملية على حساب القيمة العددية باستعمال بعض القوانين مستندة إلى أربع مزايا مهمة هي:

عدد الكلمات المفتاحية في الجملة الموجودة في السؤال بدقة.

عدد الكلمات المفتاحية الموجودة في الجملة بعد إجراء التحليل الصرفي.

تسلسل ورود الكلمات المفتاحية من الناحية الظاهريّة بين السؤال والجملة.

تسلسل ورود الكلمات المفتاحية بشكل مباشر في السؤال.

بيَّنت نتائج البحث أن نسبة الحصول على نسبة استرجاع تعادل ٩٧,٥٪ (نسبة الأجوبة المتحصل عليها بالنسبة للأجوبة ذات العلاقة الموجودة في الوثيقة) بينما بلغ مقدار دقة النتائج ٦٦,٢٥٪ (عدد الأسئلة التي كانت إجاباتها صحيحة من مجموع الأسئلة المقدمة) وهذه الإجابات أفضل من بعض النتائج التي لوحظت في مراجع أخرى.

Proc. of COPSTIC'03, Rabat, Morocco December 11-13

Year of publication: 2003

Arabic User Search Query Correction And Expansion

T. Rachidi, M. Bouzoubaa, L. Elmortaji, B. Boussouab, A.Bensaid

تصحيح استعلام بحث المستخدم العربي وتوسيعاته

إنّ محركات البحث(search engines) هي أدوات شائعة الاستخدام للوصول إلى صفحات الإنترنت، لكنها تعود بمئات النتائج والروابط التي تكون أحيانا غير ذات صلة بموضوع البحث الرئيس مما يضيّع على الباحث وقتا طويلا للوصول إلى مبتغاه، لذلك فإن استخدام أدوات البحث المتوفرة على الإنترنت بدون استراتيجية بحث محددة تشبه إلى حد كبير من يسير في مكتبة ضخمة بشكل عشوائي يحاول العثور على كتاب معين أو معلومة ما. ويعرف الاستعلام (Query) على أنّه تلك الكلمات المفتاحية التي يستخدمها الباحث للوصول إلى مستندات وروابطَ حول موضوع ما، كما يعرف توسيع الاستعلام (query expansion) على أنّه إضافة مرادفات مناسبة لكلمات الاستعلام الأصلي بشكل آلي دون تدخل المستخدم، مما يساعد في استرجاع بيانات وروابط ذات صلة.

ويُعدُّ تصحيح الاستعلام المدخل ميزة أخرى تقدمها محركات البحث، فعلى سبيل المثال عند كتابة كلمة والبحث عنها في محرك جوجل (Google) تظهر رسالة نصها «هل تقصد كذا». ويتم تصحيح الاستعلام بطرق مختلفة إما أن تكون إحصائية تعتمد على التكرار والكلمات السابقة والتالية للكلمة المفتاحية، أو تكون معتمدة على قواميسَ لغويّةٍ تحتوي بدائلَ محتملة للكلمة المدخلة.

أمّا ما يتعلق بتوسيع الاستعلام فقد تطرّقت الدراسة لعدة تقنيات بعضها يعتمد على إعادة إعطاء وزن لكلمات الاستعلام اعتماداً على التغذية الراجعة المرتبطة (relevance feedback) حيث تقوم هذه الطريقة بزيادة وزن مصطلحات الاستعلام اللاحقة التي تظهر

في الوثائق ذات الصلة، وفي خفض وزن عبارات الاستعلام التي تظهر في الوثائق غير ذات الصلة. كما توجد أيضا تقنيات مختلفة مثل تغذية المتجهات الراجعة (vector feedback) والطريقة الاحتمالية (probabilistic) لحساب الأوزان. ومن ناحية أخرى، لا يستخدم توسيع طلب البحث بدون إعادة حساب أوزان الكلمات المفتاحية، كما يمكّن توظيف الموسوعات (thesauri) لتوسيع الاستعلام الأصلي استخدامَ التعليقات ذات الصلة، واستخدام الموسوعة للتوسيع بإضافة عبارات إلى الاستعلام الأصلي. وقد بيّنت الدراسات أنّ توسيع طلب البحث حسب المصطلحات ذات الصلة أقل فاعلية من توسيع طلب البحث بحسب المصطلحات من المستندات ذات الصلة.

تناولت هذه الورقة البحثية أيضا وصفاً لأهمّ الطرق المستخدمة في محرك البحث المسمى برق (Barq) من أجل تصحيح الاستعلام المدخل من المستخدم وتوسيعه، حيث يعتمد تصحيح الاستعلام في محرّك برق على التركيب الصرفي للكلمات (word morphology) وعلى مستوى المعجم (lexicon level) دون تحليل الاستعلام المدخل من المستخدم (user query) مع الأخذ بعين الاعتبار تصحيح أخطاء الطباعة للغة العربيّة واللغة العربيّة المترجمة، لأنه مصمم خصيصاً لأغراض نطق اللغة العربيّة والأخطاء الإملائيّة، كذلك أخطاء اللغة العربيّة المترجمة وأخطاء اللغة العربيّة لمتعلميها. وعلى صعيد توسيع الاستعلام، اهتمّ محرّك برق باستخلاص جذور الكلمات لتوسيع الاستعلام بثلاث طرق مختلفة تعتمد على الموسوعات، وهي: موسوعة للمفاهيم التي تُبنى يدويّا، وموسوعة للمفاهيم التي تُبنى تلقائيّا من مستندات (XML) جرى المرور عليها (crawled documents)، وموسوعة للمفاهيم التي تُبنى تلقائيّاً من تصنيف تلقائي من المستدات التي جرى استعراضها.

Journal of Advanced Computer Science and Technology Research, Vol.3 No.1, March 2013, 1-20

Year of publication: 2013

Improving Q/A Using Arabic Wordnet

Lahsen Abouenour, Karim Bouzoubaa, Paolo Rosso

تحسين السؤال والجواب باستعمال "ووردنيت" العربيّة

مع تضخّم المحتوى المتاح على شبكة الإنترنت، أصبحت أنظمة الأسئلة والإجابة ((Q/A) Question/Answering) من أكثر أدوات البحث شهرة، حيث لاقت اهتمام عدد كبير من الباحثين والمستخدمين، وفي الآونة الأخيرة طُوِّرت العديد من هذه الأنظمة فيما يتعلق باللغة العربيّة. البحث الحالي يقدم نهجاً لتحسين نظام الأسئلة والإجابة من خلال عملية توسيع طلب البحث ((Query Expansion (QE). ويستند هذا النهج على الأنطولوجيا (ontology) التي بُنيت باستخدام قاموس وردنيت (WordNet) العربيّ. ووردنت العربيّ هو قاعدة بيانات معجميّة للغة العربيّة تابعة لمنصة الوردنت العالميّة، وتهدف إلى تجميع الكلمات العربيّة ذات المعنى الواحد ضمن مجموعات من المرادفات، بحيث يكون لكل مجموعة تعريف قصير وعام، بالإضافة إلى توضيح العلاقات الدلاليّة المتنوّعة بين المجموعات المترادفة.

في هذا البحث قام الباحثون بتصميم نظام لتوسيع الاستعلام بالاعتماد على العلاقات الدلاليّة التي تربط بين مفاهيم الأنطولوجيا. وقد أطلقوا على نظامهم اسم ((AAWN) Amine Arabic WordNet)، حيث يعرف نظام (AAWN) وهو برنامج بلغة جافا مفتوح المصدر متعدد الطبقات مخصص لتطوير الأنظمة الذكية وأنظمة متعددة الوكلاء (multi-agent systems)، ويتألف من أربع طبقات هي: طبقة الأنطولوجيا والطبقة الجبرية وطبقة البرمجة وطبقة أنظمة متعددة الوكلاء. ولبناء النموذج استخرجت كافة البيانات الموجودة في منصة (AWN) وهي مصدر مجاني للغويات العربيّة، ثم استخدمت العلاقات الموجودة بين المفردات الإنجليزيّة

بالاعتماد على وردنيت، ثم أضيفت المرادفات العربيّة لهذه الأنواع استنادا إلى علاقة التكافؤ بين المفردات الإنجليزيّة ومفردات الوردنت العربي. ثم رُبط بين هذه المفاهيم بشكل هرمي بناءً على تعريف المفهوم ومرادفاته وتمثيله في السياق. وهكذا فإنّ الشكل الهرمي يمكّن توسيع الاستعلام من تعيين وزن كلّ كلمة مفتاحية (**produced keyword**) تنتج وفقا لعلاقتها مع كلمة المصدر (**source keyword**). وهذا الوزن تعتمد وظيفته على نوع العلاقة والمسافة بين المفهوم الأوليّ والكلمة المفتاحية المنتجة.

وقد أُجريت بعض التجارب الأولية التي تبيّن تحسن إمكانية الحصول على الجواب المتوقع في الوثائق التي أُرجعت عند استخدام النهج المقترح. وهذه التجارب أظهرت قدرة النظام على الحصول على بعض الإجابات في حين فشل جوجل في إيجادها.

٣-٤-٨ أبحاث تلخيص النصوص

وتضم ثمانية أبحاث بينها بحث مسحي واحد بعنوان: التّلخيص التلقائي للنّص العربي: دِراسة مَسحيّة، وبحث واحد من نوع (أ) بعنوان: الاستخراج التّلقائي للعلاقات الأنطولوجية من النّص العربي.

وهناك ستة أبحاث من نوع (ب)، هي: تلخيص النّص العربي القائم على نظريّة المخطّطات، ونموذج لتلخيص النّص العربي باستخدام تقنيات التّجميع العُنقودي، و تلخيص النص العربي التلقائي باستخدام العنقدة واستخراج العبارة الرئيسية، و مُلَخِّص (برنامج تلخيص) قائم على استخراج العبارات الرئيسيّة، والتكرار الأقل والارتباط الأكبر لتلخيص ملفات النصوص العربية المفردة والمتعددة، ونحو بناء قائمة بيانات قياسية لتقييم استخلاص العبارة الرئيسيّة في النّص العربي.

Artificial Intelligence Review, February 2016, Volume 45, Issue 2, pp 203–234 (Artif Intell Rev (2016) 45: 203)

Year of publication: 2016

Automatic Arabic text summarization: a survey

Asma Bader Al-Saleh, Mohamed El Bachir Menai

التّلخيص التلقائي للنّص العربي: دِراسة مَسحيّة

مقدمة

أصبح من الشائع استخدام الملخّصات، التي باتت تُستخدم كثيراً في الحياة اليومية. ومن الأمثلة عليها مُلخّصات الأوراق، وعناوين الأخبار التّلفزيونية، واستعراض (عرض) الكتب، وأدلة التسوّق الخاصة بالمتاجر، وحتى الإعلانات التقديمية (الدعائية) للأفلام. وبوجه عام، يمكن تعريف الملخص بأنه «نص ينتج من نص واحد أو أكثر، ينقل معلومات مهمة موجودة في النص الأصلي (النصوص الأصلية)، ولا يزيد عن نصف النص الأصلي أو النصوص الأصلية.

مَدارُ هذا المَسْح حولَ العديدِ من الدّراسات البَحثيّة التي أُجريت في مجال تلخيص النّص العربي، وبالتحديد ما يتعلّق بأساليب التّلخيص والتقييم، فضلاً عن الذخائر اللغوية المستخدمة في تلك الدراسات. غير أنّ الأدبيّات المستخدمة في هذا المجال محدودة وجديدة نسبياً بالمقارنة مع الأدبيّات المتاحة للُّغات الأخرى، مثل اللغة الإنجليزية. لذلك، هناك فرصة كبيرة متاحة لإجراء مزيد من البحث في مجال تلخيص النّص العربي. وبالإضافة إلى ذلك، كان من أهم المشاكل التي واجهت تلخيص العربيّة غياب الملخّصات العربيّة القياسية الذهبية، وعلى الرغم من أنّ هذا الوضع قد بدأ في التغيّر، خاصة مع إدراج اللغة العربيّة جُزءاً من الذخائر اللغوية والمهام التي نوقشت في ورشتي عمل (MultiLing TAC ٢٠١١ MultiLing Pilot and ACL ٢٠١٣). وأخيراً، يُعدُّ توفير الذخائر اللغوية المطلوبة واعتمادها في دراسات التلخيص العربيّة مطلباً أساسياً.

ونظراً للكَمّ الهائل من البيانات التي جرى ضخّها على الشّابكة، أصبح الوُصول إلى عدد هائل من موارد النُصوص مُتاحاً بسُهولة لأيّ مُستخدِم. وبالتّالي، أصبَح تلخيص النّص التِلقائي حاجة أساسية للتغلُب على مُشكلة الحُمولة الزائدة للمَعلومات، حيث يختصر التلخيص من الوقت والجُهد اللاّزمين لاستكشاف الأجزاء الأبرز والأكثر أهمية في مجموعة من النُصوص وتحديدها.

وفي الواقع، لا يُوجد تَصنيف مُحدّد لأنواع الملخصات، حيث يتفاوت تصنيف الملخّصات بحسب زاوية النظر. فعلى سبيل المثال، استناداً إلى العلاقة مع مَعْلَم النّص الأصلي، يُصنّف التّلخيص إلى تلخيص استخلاصي (اقتباسي)؛ (**Extractive summarization**) يعتمد على الجَمْع بين الأجزاء الأكثر أهميّة من المَصدر حَرْفياً دون تعديل النّص المحدد. وعلى النقيض من ذلك، هُناك التّلخيص الإجمالي (الإيجازي) (**abstractive summarization**) الذي يعني إعادة صياغة الأفكار المهمة في الوثائق الأصلية لإنتاج مُلخّصات ذات قدرٍ أكبر من القوّة النّحْوية والتّماسك. بالإضافة إلى ذلك، يُمكن لعملية التّلخيص أن تُنتج مُلخصات عامة (**generic summaries**) استناداً إلى معلمات (متغيرات) الجمهور (**audience parameters**) ، وذلك في حال إذا كانت تستخدم – فقط – المستند الأصلي، أو الملخّصات المستندة إلى الاستعلام (**query-driven**) (بدافع الموضوع) إذا كان تركيزها على استرجاع الموضوعات المهمّة المرتبطة بطلب بحث المستخْدِم. وهناك أيضاً تصنيف آخر للتّلخيص بحسب معلمة (مُتغيّر) الاتساع أو الشُمول، بحيث تُصنّف عملية التلخيص إلى تلخيص (الوثيقة الواحدة) إذا كان الملخص مبنيا على أساس وثيقة واحدة فقط أو (التلخيص متعدد الوثائق) إذا كان الملخّص مبنياً على أساس وثائق مُتعددة. وهناك أيضاً التصنيف حسب مَعْلمة اللُّغة، وتُعدُّ أيضاً مُهمة في تصنيف الملخصات إلى: (أحادية اللُّغة) أو (ثنائية اللُّغة) أو (مُتداخلة–اللُّغات).

وأخيراً، خلافاً للدّراسات المرجعيّة والمسحيّة الأخرى حول التلخيص التلقائي، يركز هذا المسح على مسألة تلخيص النص العربي. ويهدف إلى تقديم رقم عام حول واقع التلخيص التلقائي للنص العربي من خلال استقصاء العديد من الدراسات البحثية الموجودة في هذا المجال،

حيث ناقشت الدراسة أساليب تلخيص النص العربي وتقييم النتائج وكذلك عرض الذخائر اللغوية ذات الصلة والمؤتمرات وورش العمل التي عُقِدت في هذا المجال. واشتملت الورقة على أربعة مباحث رئيسية هي: الخصائص البارزة للغة العربيّة، وتقييم التلخيص، وأساليب التلخيص والمناقشة، وأخيرا: الاستنتاجات والتّوصيات الرئيسية للبحوث المستقبلية.

مشكلة البَحْث

رغم مرور ما يزيد عن ٥٠ سنة على هذا المجال (التّلخيص التلقائي)، لا تزال هناك العديد من المشاكل المحيطة به، بحاجة إلى حُلول. فعلى سبيل المثال، تُعدُّ إدارة التحدّيات التي تواجه تقييم التلخيص مسألة مفتوحة في تلخيص النُصوص. وفي الواقع، يُعدّ التّلخيص التلقائي مجالاً نشطاً، خاصة فيما يتعلق باللغة العربيّة، حيث تُعد الأدبيات التي تناولت التلخيص التلقائي للنّص العربي ضئيلة وحديثة نسبياً، وتنمو ببطء، مقارنة بحجم الأدبيات بالنّسبة لِلُغات أخرى مثل اللغة الإنجليزية.

المبحث الأول: الخصائص البارزة للُغة العربية

تُعد اللُّغة العربيّة اللغةَ الرسمية لـ ٢٢ دولة، ويتحدّث بها أكثر من ٣٠٠ مليون شخص حول العالم، وهي إحدى ست لغات رسمية مُعتمدة في الأمم المتحدة. وينطوي التّعامل مع اللُّغة العَربيّة على عدّة تحدّيات بالنّسبة للعاملين في مجال «مُعالجة اللغة العربيّة الطبيعية»، التي من شأنها أن تُعقّد المهام المختلفة مثل تحديد جذر الكلمة، والترجمة الآلية، والتلخيص التلقائي، وحتّى تقطيع الكلمة إلى حروف. ومن ضمن هذه التّحديات ازدواجية اللّسان العربي، ووجود درجة عالية من الغُموض بالإضافة إلى الطّبيعة الاشتقاقيّة والإعرابيّة لهذه اللغة. ويستعرض هذا المبحث -فيما يلي- بعضاً من هذه التّحديات مع تقديم أمثلة تفسيريّة عليها.

كما أن هناك عدة لهجات عربيّة، حتى ضمن البلد العربي الواحد، لكن الشعوب النّاطقة بالعربيّة تستخدم اللّغة العربيّة التقليدية في الصّلوات اليوميّة، في حين تُستخدم اللُّغة «العربيّة

الموحدة الحَديثة» في الأنواع المختلفة من الخِطابات ووسائل الإعلام كالصُحُف والمجلات ومحطات إذاعة الأخبار، ولذلك جاء تركيز مُعظم الجُهود المتعلقة بدراسة «مُعالجة اللُّغة العربيّة الطّبيعية» على «العربيّة الموحدة الحديثة»، رغم الحاجة لدراسة أنواعٍ أخرى من اللُّغة العربيّة إلى جانب القواعد النّحوية والمعاجم (المفردات) وخاصةً في سياق شبكات التّواصل الاجتماعي. وهُناك عدة عوامل تُساهم في مُشكلة الغموض، وأحد هذه العوامل غياب حُروف العلّة القصيرة التي حلّت محلّها علامات التّشكيل التي غالباً ما يتمّ إهمالُها في النّص المكتوب لـلغة «العربيّة الموحدة الحديثة». وفي الواقع، يبدو أنّ العَديد من الدّراسات البَحثية – المتخصّصة في مجال تشكيل النّص العربي– تمكنّت من التغلّب على هذا الغُموض عن طريق إضافة التّشكيل المفقود إلى النّص العربي. وهناك أيضاً عامل آخر يُساهم في مُشكلة غموض اللُّغة العربيّة، وهو أن اللُّغة العربيّة تميل لإحلال الضّمير محل الفاعل في الجملة. ومن جهة أخرى، لا تتضمن اللُّغة العربيّة حروفاً كبيرة، الأمر الذي يُعقد مهمة تحديد الأسماء أو الاختصارات أو العناوين.

علاوة على ما سبق، تُعد اللُّغة العربيّة لغة اشتقاقية وإعْرابيّة. وهذا من شأنه أن يُضيف المزيد من التّحديات لمهام مُعالجة اللُّغة العَربيّة الطّبيعيّة، مثل تجريد الكلمة من الزوائد (Lemmatization) و تحديد جذور الكلمات (Stemming). وتمتلك اللغة العربيّة حوالي ١٠٠٠٠ جذر ونحو ١٢٠ نمطاً لإضافة لاحقات.

وأخيراً، هناك مسألة أخرى تتعلق بضبط الكتابة أو الإملاء باللُّغة العربيّة، وتُساهم أيضاً في مشكلة الغموض من خلال إهمال كتابة العلامات، مثل كتابة النّقاط أو الهمزات على الحروف. ونتيجة لذلك، يمكن كتابة عدة كلمات مختلفة بنفس الطريقة. لذلك، كما هو الحال مع جميع تطبيقات معالجة اللغة العربيّة الطبيعيّة، يجب التعامل مع النص العربي بعناية فائقة خلال عملية التلخيص للتغلُّب على التّحديات المذكورة أعلاه. لكن لسوء الحظ، هناك عدد محدود من الأدوات والموارد العربيّة المتاحة لجمهور الباحثين.

المبحث الثاني: تقييم التلخيص

هناك عدة عوامل تقف خلف التحديات التي تواجه مسألة تقييم التّلخيص والصعوبات ذات الصلة بإجراء مقارنة بين أنظمة التلخيص المختلفة. وأحد هذه العوامل أو الأسباب هو أنّه من أجل تقييم محتويات الملخّص يجب إجراء مقارنة بواسطة ملخصات نموذجيّة أو مرجعيّة اعتماداً على ذخيرة لغويّة قياسية تحتوي على كل من الوثائق الأصلية المراد تلخيصها والملخصات المرجعيّة.

ويُناقش هذا المبحث فيما يلي المؤتمرات والورشات والذخائر اللغويّة ذات الصّلة بتلخيص النص العربي من جهة، وطرائق التقييم المستخدمة في الأدبيات التي شملها هذا المسح.

أولاً: المُؤتمرات وورشات العمل والذخائر اللغويّة

منذ العام ١٩٩٨، نُظّمت العديد من المؤتَمرات وورش العمل التي ناقشت موضوع التّلخيص التلقائي مثل: **SUMMAC** و **DUC** و **TAC** وقد أفادت هذه المؤتمرات حقل التلخيص بطُرق عدة، من خلال تزويد الباحثين بالذخائر اللغوية المطلوبة وتمكينهم من التقييم والمقارنة بين طُرقه المختلفة.

وكما أسلفنا في المقدمة، فقد أُدرجت اللغة العربيّة في ورشتي عمل (**MultiLing 2011 Pilot and MultiLing 2013**). حيث ناقشت الورشة الأولى موضوع التقييم والتلخيص مُتعدد اللُّغات، بالإضافة إلى مسائل جمع البيانات، وتمّ الاعتماد على ذخيرة لغويّة تجريبية مُتعدد اللغات (**TAC 2011 MultiLing Pilot 2011**)، وهي عبارة عن ذخيرة لغويّة لملخّصات مُتعدّدة اللغات ومُتعددة الوثائق تحتوي على مجموعات من الوثائق المكتوبة بسبعِ لغات، من ضمنها اللُّغة العَربيّة. وهُناك العَديد من الدّراسات التي حاولت مُعالجة مشكلة النقص في ذخائر لغوية عربيّة.

ثانياً: طرائق التقييم

يستعرض هذا القسم طرائق التقييم المختلفة المستخدمة في الدراسات التي شملها هذا

المسح، حيث قدّم مُلّخصاً لأنواع التّقييم (اليدوي والآلي) والأساليب المستخدمة. على سبيل المثال، استُخدم التقييم البشري في بعض هذه الدراسات بالإضافة إلى أسلوبي الدقة والتذكير، التي تُعد من المقاييس الشائعة لاسترجاع المعلومات التي استخدمت من عدة دراسات، كما استُخدمت عدة طُرق تِلقائية أُخرى (مثل MeMoG variation of AutoSummENG) في ورشة عمل (MultiLing TAC 2011 MultiLing Pilot and ACL 201) لتقييم الأنظمة المُستخدمة في ورشة عمل (TAC 2011 MultiLing Pilot).

المبحث الثالث: أساليب التلخيص

كانت أولى محاولات دراسة تلخيص النصوص قد نُشرت قبل أكثر من خمسين عاماً وتضمّنت مجموعة مُلخّصات أُخذت من وثيقة واحدة فقط. ثم بدأ استخدام أسلوب البحث عن تكرار الكلمة أو المفردة لتحديد أهمية كل جُملة. ولاحِقاً، استُخدمت سمات أكثر (مثل تكرارات الكلمة ومواضع الجمل والكلمات الجديلة (التأشيرية) (cue words)، جُمعت خطّياً لحساب علامات الجُمل، وما زال بعضٌ منها مستخدماً حتى اليوم. وقد صُنِّفت في هذا المبحث الدراسات التي شملها المسح وفقاً لطريقة التلخيص السائدة على أساس التصنيف المذكور أعلاه لأساليب التلخيص.

الأسلوب الرمزي (Symbolic approach): وهو أسلوب يعتمد على عدة خطوات لعمليّة التّلخيص: ابتداءً بتحديد الوحدات النّصية (textual units) على أساس العبارات التأشيرية (cue phrases) ثمّ تكوين علاقة بلاغية بناءً على هذه العبارات، ولاحقاً استخدام هذه العلاقات لبناء ما يعرف بأشجار RS-trees ، وأخيراً، اختيار أفضل شجرة لتكوين الملخص.

الأسلوب العددي (Numerical approach): كانت أولى محاولات وضع ملخِّص (أداة تلخيص)(summarizer) للنص العربي عام ٢٠٠٤ (Douzidia & Lapalme)، من خلال ابتكار نظام تلخيص للنّص العربي أُطلق عليه اسم (Lakhas) وكان الهدف منه تقديم ملخص قصير جداً (أي عنوان) مكوّن من عشر كلمات تقريباً لتلبية متطلبات المهمة الثالثة للـذخيرة اللغويّة (DUC ٢٠٠٤) وتضمن وثائق بالعربيّة وترجمتها إلى الإنجليزية. وكان أحد أهم الدوافع

الرئيسية لهذه الدراسة هو الحد من تأثير أخطاء الترجمة الآلية وذلك من تلخيص الوثائق العربيّة مباشرة بدلاً من تلخيص الوثائق المترجمة. وقد صُمم هذا النظام ليعمل بواسطة اختيار جُمل الملخّص اعتماداً على درجاتها (علامتها)، ويُحتسب ذلك باستخدام توليفة خطّية مكوّنة من أربع سمات، أما السمات الثلاث الأولى فهي (الصّدر، العنوان، الإشارة) في حين تتعلق السّمة الرابعة بقيمة (TF-IDF) وقد أُعطي لكل منها قيمة معينة بحسب موقع الجُملة. ثمّ ظهرت لاحقاً محاولات عدة لاقتراح مُلخِّصات (أدوات تلخيص) تلقائيّة اعتماداً على تكرار الكلمات أو اقتفاء جذور الكلمات أو اعتماداً على التحليل الدلالي للنصوص. كما اقتُرحت أنظمة تلخيص (عربي-إنجليزي) تلقائيّة مُتعددة الوثائق، ونظراً لغياب ملخصات موحّدة يُعتمد عليها في ذلك الحين، كان التركيز على الذخائر اللغويّة. إضافة لما سبق، ناقش هذا المبحث نتائج العديد من الدراسات المقارنة بين الجُهد الآلي أو البشري في التلخيص والتقييم واعتماد التلخيص على الترجمة الآلية او الترجمة البشرية. كما تضمنت العديد من الدراسات حول تلخيص النص العربي إدخال تقنيات التعلم الآلي في الملخِّصَات المقترحة.

الأسلوب الهجين (hybrid)

هناك دراسات أخرى شملها هذا المسح اعتمدت الأسلوب الهجين لبناء نظام تلخيص عام (استخلاصي) لإنتاج مُلخصات بأحجام مختلفة بناءً على خيار المستخدم، بحيث يُنتج الملخص في ثلاث خطوات رئيسية: أولا: تطبيق أسلوب الشجرة (RS-tree) لإنتاج النُسخة الأولى (المبدئيّة) من الملخص؛ ثانيا: إعطاء كل جملة في الملخص المبدئي النتيجة (العلامة) التي أمكن الحصول عليها عن طريق جمع خمس سمات مُستخرَجة من هذه الجمل. وهي موضع الجملة، والتكرارات الإجمالية لكلماتها، وهل لها أرقام، وهل تقع في السطر الأول من الوثيقة، وهل تحتوي على كلمات من العنوان؟. وأخيراً: اختيار الجُمل التي حصلت على أعلى علامة.

وفي دراسة أخرى اتّبعت الأسلوب الهجين، أُجريت مقارنة نتائج أسلوبين لاستخلاص الفقرات المهمة من الوثائق العربيّة لإجراء التلخيص، الأول: بناء التمثيل البلاغي للنّص

باستخدام أسلوب الشجرة، والثاني: يمثّل النص الذي يستخدم التمثيل اعتماداً على (vector space model (VSM)).

واستعرض هذا القسم أيضاً دراسة حالة لنظام تلخيص تلقائي للغة العربيّة (كتبت وثيقته الأصلية باللغة الفرنسية) حيث يقوم النظام باستخراج الجمل التلخيصية استنادا إلى أسلوب هجين يجمع ما بين التحليل الرمزي (البلاغي) (symbolic (rhetorical analysis)) والتقنيات العددية (التعلم الآلي)(numerical (machine learning)). وكانت آخر الدراسات التي ناقشها هذا القسم حول الأسلوب الهجين المكوّن من طريقتي التماسك المعجمي وتجزئة النص العربي في التلخيص.

المبحث الرابع: المناقشة

لخّصت الدراسة ضمن هذا المبحث نتائج الدراسات والأدبيات التي شملها هذا المسح مثل: طرائق التلخيص والسمات المستخلصة والذخائر اللغويّة المستخدمة. وأبرز هذا المبحث أهمّ الملاحظات التي توصّلت اليها تلك الدراسات حول هذا الموضوع.

ومن الواضح – حسب المبحث الثالث – قلة الدراسات التي تناولت موضوع تلخيص النص العربي مُقارنة باللُّغات الأخرى. وأخيراً، فيما يتعلّق بتقييم التّلخيص، الذي يهدف إلى التعرّف على مُستوى الأداء ومقارنة نتائج أنظمة التّلخيص المختلفة، فهو يُمثّل إحدى الصُعوبات في مجال التّلخيص التلقائي. وفي الدراسات التي شملها المسح، استخدم التّقييم البشري بالإضافة إلى التقييم الآلي، لكن رغم اعتماد العديد من الأنظمة على التقييم البشري ونجاحه في الارتباط بتلك الأنظمة المعتمدة على الذخائر اللغويّة الإنجليزيّة، فإنّه ليس بالضرورة أن تَنْجح هذه الفرضيّة مع الأساليب الأخرى للتقييم البَشري.

الخاتمة والتوصيات

تضمّنت هذه الدراسة المسحيّة عرضاً للأدبيات ذات الصّلة بالتّلخيص التلقائي للنّص العربي، حيث ناقشت الدراسة الطُرق المختلفة المستخدمة في إنتاج الملخّصات أو تقييم نتائج

عمليات التّلخيص. وأظهرت نتائج المسح قلة الدراسات والأبحاث الحديثة في هذا المجال. وعليه، برزت العديد من الاتجاهات البَحثيّة الداعية لتقصّي إمكانية تحسين عملية التّلخيص بحدّ ذاتها والجوانب الأساسية المرتبطة بها مثل: التّرجمة الآلية وأساليب التقييم والمعالجة المسبقة، الخ.

أضف إلى ما سبق، ضرورة استكشاف النّهج الدلالي وإنتاج مُلخّصات أكثر اختصاراً، إلى جانب أنواع التلخيص الجديدة. ويجب أن تستكشف دراسات تلخيص النص العربي ثراء أساليب التّلخيص في الأدب الإنجليزي. كما أنّ مهام تلخيص النص العربي يمكنها الإفادة من الدراسات الأخرى في مجال مُعالجة اللغة العربيّة الطبيعيّة.

ووفقاً لهذه الدراسة، يبدو أنّ إحدى المشاكل الرئيسية في تلخيص النص العربي هي غياب استخدام مُلخّصات عربيّة ذهبية معتمدة يمكن الاستناد إليها. ويُمكن التغلّب على هذه المشكلة من خلال الجُهود المختلفة التي بذلت لإنشاء ذخائر لغوية عربيّة، لا سيّما في ورشتي العمل (TAC 2011 MultiLing Pilot and MultiLing 2013) اللتين شجّعتا التطبيقات المختلفة للمُلخّصات المستقلة- لغويّاً بلُغاتٍ مُختلفة، من ضِمنها اللغة العربيّة.

وفي الختام، إذا استُخدِمت هَذه الذخائر اللغوية واعتُمِدت في الدّراسات المستقبليّة، فإنّها قد تُعْتمَد مُلخّصاتٍ عَربيةً قياسيّةً ذهبيّةً. وأخيراً، يُعد توفير الموارد والذخائر اللغويّة حاجةً مُلحّة من شأنها أن تُشجّع آفاق بحثٍ جديدة من خلال تزويد الباحثين بأساسٍ متين لمقارنة نتائجِهم.

Journal of King Saud University – Computer and Information Sciences (2014) 26, 462–472

Year of publication: 2014

Automatic extraction of ontological relations from Arabic text

Mohammed G.H. Al Zamil, Qasem Al-Radaideh

الاستخراج التّلقائي للعَلاقات الأنطولوجية من النّص العربي

المقدمة

يُعرّف مصطلح الأنطولوجيا بأنّه ()علم توصيف المفاهيم"، وهي النّمذَجة الاصطلاحية للمُكوّن اللُّغوي على امتداد علاقاته الدلالية فيما يتعلق بالمفاهيم الأخرى. ويمكن النظر إلى الأنطولوجيا باعتبارها نمطا لكيفية تصميم مفهوم معين ليكون مُرتبطاً بالمفاهيم الأخرى الموجودة في سِياق مُعين.

ويُعد الاستخلاص التلقائي للعلاقات الدلالية بين المفاهيم العربيّة لغرض صياغة نماذج الأنطولوجيا أمراً بالغَ الأهمية لتوفير بيانات شرحيّة دلالية غنيّة. وبسبب الزّيادة السّنوية للمُحتوى العربي على الشّابكة، ظهرت الحاجة إلى أدوات مُتخصّصة لتحليل النص العربي وفهمه. ويقترح هذا البحث منهجية لاستخلاص العلاقات الأنطولوجية، ويهدف إلى: استخراج السِمات الدلالية للنص العربي، واقتراح أنماط نحْوية للعلاقات بين المفاهيم، واقتراح نموذج اصطلاحي لاستخراج العلاقات الأنطولوجية.

وصُمّمت المنهجية المقترحة لتحليل النّص العربي باستخدام الأنماط الدلاليّة المعجميّة للغة العربيّة وفقاً لمجموعة من السّمات. بعد ذلك، جرى تلخيص السّمات وإثراؤها بأوصاف اصطلاحية لغرض تعميم القواعد الناتجة. ولاحقاً، صيغ مُحلِّل القواعد اللُّغوية الذي يتقبّل

النّص العربي ويُحلِله، ثم يعرض المفاهيم ذات الصّلة الموسومة بعلاقاتها المميزة لها. ولغرض حلّ الغموض الذي يكتنف ألفاظ الجناس، جرى تكرار استخدام قائمة من أدوات الترجمة الآلية، والتنقيب عن النّص، وبعض من خوارزميات وَسْم الكَلام وإعادة استخدامها. كما اعتمدت هذه الدراسة على إجراء تجارب موسّعة لقياس فعاليّة الأدوات المقترحة. وتشير النتائج إلى أنّ المنهجيّة المقترحة واعدة لأتمتة عملية استخراج العلاقات الأنطولوجية.

مُشكلة البَحث

إنّ تطوير أنطولوجيا من النّص العربي عملية مُعقّدة، لأنّ استخراج العلاقات الدلالية بين المكونات اللُّغوية ما زال يعتمد على البنية النحوية (الإعرابيّة) للُّغة. ومع ذلك، فإنّ تفسير النص المستقل يتطلب تحديد نوع المعلومات التي ستُعالج والطريقة التي سيُعبّر بها عنها. وعِوضاً عن شرح كل شيء في النص (أي التّحليل النحوي)، يمكن البحث فقط عن العلاقات المعجميّة المعروفة. وهكذا، يمكن العُثور على معلومات ذات معنى باستخدام خوارزميّات بسيطة وسلسة، الأمر الذي يُؤدي إلى أتْمتة سَلِسة للعملية.

منهجيّة البحث

لتطبيق خوارزميّة هيرست (Hearst) لاستخراج العلاقات الأنطولوجية من الذخيرة اللغوية العربيّة، هناك حاجة إلى مزيد من التعزيزات للتكيّف معها. في هذا القسم، سنُلّط الضّوء على التقنية المقترحة التي اعتُمد عليها لتعديل خوارزمية هيرست (Hearst)، ودمجها ضمن الإطار العملي، الذي يتألف من خمسة مُكوّنات وظيفيّة (functional components). ومن النّاحية الاصطلاحية، ومن أجل حَلّ غموض الجِناس (ambiguity of having homonyms) أو المفاهيم التي تشير إلى سِياقات مُختلفة، قُسّمت الأنماط الدلالية إلى جُزأين (إيجابي وسلبي). يُمثل الجُزء الإيجابي وجود علاقة صحيحة، في حين يمثل الجُزء السّلبي المفهوم غير ذي الصّلة.

المُعالجة المُسبقة واستخراج السمات

تؤدي مهام المعالجة المسبقة واستخراج السمات دوراً مهماً في تسهيل مُعالجة النص أثناء التّحليل، لكن من الضّروري أولاً وصف النص المدخل بشكله الجَديد. كما أنّ تحديد السمات التي ستُستخدم في الكشف عن الأنماط النصية وتبريرها أمرٌ ضروري ويؤثر على دِقّة المنهجية المقترحة بعُمومها. وتُعرف السّمات، في هذا السياق، بأنها «مكوّنات اللغة والعلاقات الموجودة بينها». وفي حين يُفترض أنّ استخلاص المكونات اللُّغويّة يجري تلقائياً، فإنّ العلاقات، بالمقابل، تُحدّد يدوياً للحدّ من نطاق التّجريب. وقد استُخرج خلال هذه الخطوة أربع سمات مُختلفة يُعتقد أنّها تفي بمُتطلّبات بناء الأنماط النّحْويّة المعجمية للنّص العربي. وقد أوضحت العلاقات التي تتعيّن دراستها في هذا البحث ووصف هيكلها كمجموعة من الأمثلة. تدلّ العلاقة بين العام والخاص (الجناس) على أنّ هناك علاقة دلاليّة بين مفهومين: الخاص (أو الجناس Hyponym) والعام (Hypernym). وهي العلاقة بين السبب و المسبّب التي تحدد نماذج العلاقة السببيّة، وتجري من خلالها نمذجة السبب والنتيجة لبعض الأفعال. بالإضافة إلى علاقة الجُزء-الكُلّ، و علاقة (Is-a)،(Has-a) ونوع نموذج العلاقات الهرميّة بين المفاهيم.

الأنماط النّحْوية المُعجميّة

في هذا القسم، وُصفت النُسخة المحسّنة من خوارزمية هيرست المطبقة على النّص العربي، وأوضحت المهام المطلوبة لتطبيق الخوارزمية، حيث نُفّذت الخطوة الأولى يدوياً؛ إذْ كان يجب على الخُبراء اللغويين المختصّين في مجال معين من المعرفة أن يُقرروا العلاقات المناسبة، ثم أَمكن استخراج أمثلة بشكل تلقائي لتشكيل قائمة التمرينات. وبعد توفير العلاقات وقوائم الأمثلة التدريبية، شُغلت الخوارزمية للعُثور على حالات مماثلة في النص، وبمجرد توفر الحالات الموجودة تفترض الخوارزمية أن تلك الحالات بنية عامة تقوم على تعميم النمط باستخدام علامات (وسم) (tagging) أقسام الكلام. لكن يجري إخضاع جميع الأنماط المستخرجة لمهمة التقييم التي تختبر شمولية كلّ نمط لإزالة الشوائب منها، وتغطي حالات قليلة جداً أو حتى لا تكاد تُذكر.

ولتسهيل تنفيذ الخوارزميّة على النص العربي الذي قد يحمل قدراً كبيراً من العَلاقات المعقّدة (العام-الخاص)، تتطلب خوارزميّة هيرست بعض التعديلات للإشراف على جُزء منها، مما يؤدي إلى إنتاج منهجية شبه مُراقَبة. وقد استُوفي هذا الشرط عن طريق تحويل كل نمط إلى استعلام يُستحضر لغرض إيجاد نص مطابق.

مَرحلة التوسّع

ولتجنّب وجود أنماط زائدة أو مكرّرة (**redundant patterns**) تشير إلى نفس المفاهيم، وُسّعت مرحلة تقييم البنية المعجمية للنماذج لتشمل المرادِفات. وفي الواقع، نعتقد بأنّ هذه المرحلة سوف تحسّن العدد الإجمالي للأنماط المستخرجة، كما أنّ هذا التوسّع من شأنه أن يؤدّي إلى استخلاص علاقات جديدة بين المفاهيم التي لم تظهر بوضوح في النص. ولتحقيق هذا الهدف، قمنا بإنشاء مقطع من برنامج يستدعي المترادفات المتكررة في أداة الشّبكة العَربيّة العالميّة.

وعلى الرغم من أنّ هذا التوسّع ساهم في اكتشاف علاقات جديدة، فإنّ بعض الأنماط القائمة أصبحت زائدة عن الحاجة، أي أنّها تُعطي نفس التأثير على مجموعة البيانات. ولتجنب التكرار، استُخدم أسلوب آخر جديد للترشيح (التّصفية) للكشف عن الأنماط الزائدة (المتكررة) وإزالتها.

تَصفية الأنماط وتجميعُها

وبما أنّ الطّريقة المقترحة شبه مُراقبة، وبسبب تأثير مهمة التوسّع على القائمة الناتجة، فإنّ بعض الأنماط الناتجة قد تغطي نفس العلاقة. وللتغلّب على مثلّ هذه المشاكل (أي العلاقات العابرة، والمرادفات، وتمثيل المفهوم)، طبقنا مقياس التغطية الذي يعمل من خلال تحديد نطاق تغطية نمط ما ضمن قائمة بيانات مُحددة. فإذا كان أحد الأنماط يُغطّي نفس أمثلة البيانات لنمط آخر، يتم إزالة الثاني منهما. وبالتالي، فإنّ تطبيق هذه القواعد باستخدام خوارزمية تحقُق جيدة (**well-formed validation algorithm**) سوف يؤدي إلى التقليل من الأنماط الناتجة، وهذا بدوره يحسّن الأداء العام للإطار المقترح.

النتائج

يقدم هذا القسم وصفاً تفصيلياً للتجارب التي أُجريت على قوائم البيانات المختلفة، والنتائج التي استُحصلت من حيث الدّقة و الاستدعاء، واختبار مقياس-F. كما أُجري تحليل الحساسية (sensitivity analysis) الذي يبيّن أثر الظواهر المختلفة على مختلف مقاييس الأداء. علاوة على ذلك، أُجريت مقارنة بين تلك النتائج ونتائج التقنيات المماثلة على قوائم البيانات العربيّة. وأخيراً، سُلّط الضوء على الأخطاء الرئيسيّة التي واجهت تطبيقنا للتقنية المقترحة خلال التجارب، مع ملاحظة النتائج فُسّرت في جزء المناقشة لتعليل مدى متانة عملنا المقترح أو ضَعفه.

مناقشة النتائج

أظهرت النتائج التجريبيّة أن الإطار المقترح يتّسم بالكفاءة في حال جرى تنفيذُه لتوليد العلاقات الأنطولوجية ضمن المفاهيم العربيّة. وخلال هذه التجارب، وجدنا أنّ أداتنا المقترحة حصلت على ٤٧ ,٧٩٪ عند تطبيقها على قائمة البيانات الخاصة بالقرآن الكريم، و٠٥ ,٨٧٪ على قائمة البيانات الخاصة بالصُحف، و ٠٩ ,٧٢٪ على بيانات المدونات من حيث اختبار- F. و نعتقد أنّ تعزيز الخوارزمية المقترحة من خلال مرحلة التوسّع، ومهام الترشيح (التصفية) (filtering task) والتحقق (validation task) وتطبيق الأنماط السلبية (negative patterns) سوف يؤدي دوراً مهما في تحسين مستوى الدقة.

الخاتمة

استعرضت هذه الورقة البحثية الأسلوب التلقائي لاستخراج العلاقات الأنطولوجية من النّص العربي، واعتمدت التّقنية المُقترحة على تنفيذ نُسخة مُحسّنة من خوارزمية هيرست (Hearst)، وأُجريت تجارب مكثّفة لقِياس أداء الطّريقة المقترحة، ودراسة تأثير العوامل المختلفة على أداء التقنية المقترحة، وإجراء مقارنة مع الطرق المماثلة. وأشارت النتائج إلى أنّ

التقنية المقترحة مُرشّحة لتمثل أسلوباً جيداً لاستخراج العلاقات الأنطولوجية من النّص العربي مقارنة مع التقنيات المتوفرة.

2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)

Year of publication: 2015

Arabic Text Summarization based on Graph Theory

Nabil Alami, Mohammed Meknassi, Said ALAOUI OUATIK and NourEddine ENNAHNAHI

تَلخيص النّص العربي القائِم على نَظريّة المُخطّطات

المُقدّمة

التّلخيص التلقائي للنّص هو عملية تهدف لاختصار طُول المستند الأصلي دون التّأثير على المحتوى من خلال استخراج معلومات مهمّة مأخوذة من كمية ضَخمة من البيانات النّصية. والهدف الرئيسي من هذه العملية هو تسهيل مُهمّة القراءة والبحث عن المعلومات ضمن الوثائق الكبيرة. وتُقدّم هذه الورقة طريقة جديدة لتلخيص النّص العربي، قائمة على أساس نَظريّة المخطّطات والتشابُه الدلالي (**semantic similarity**) بين الجُمل، لحساب أهميّة كل جُملة ضمن الوثيقة واستخراج الجُمل الأكثر أهميّة لتكوين مُلخّص الوثيقة. أضف إلى ذلك أن الكلمات التي تشترك بنفس الجَذر تكون ذات صِلة دلاليّة، كما أنّ تقنيات اختيار السّمات على أساس الجذر يُمكن أن يُحسّن من مُستوى التّشابه الدلالي بين الجُمل وأن يزيد من وزن المِيزة أو السّمة الدلالية في الجُملة. وتضمّنت المباحث الرئيسية لهذه الورقة أولاً مراجعة الأعمال ذات الصِلة في هذا المجال، وخاصة في مجال تلخيص النّص العربي، ثم استعرضت بنية النظام وتركيبه المقترح، ومكوناته، وميزاته. وأخيراً قدّمت الورقة تقييماً للنظام ومقارنته بالطُرق الأخرى الموجودة.

المُبررات ومشكلة البحث

يُعدّ تلخيص النّص أصعب مهمة في سياق استرجاع المعلومات، خاصة عندما يتعلّق الأمر باللُّغة العربيّة. وعلى عكس اللُّغات الإنجليزية والأوروبية، هناك القليل جداً من البُحوث في مجال تلخيص النّص العربي التي ما تزال في بداياتها. وما تزال أنظمة التلخيص باللُّغة العربيّة غير ناضجة وغير موثوق بها مثل تلك التي طُوِّرت باللغات الأخرى كما هي في اللُّغة الإنجليزية.

الاستنتاجات والتوصيات

يُقدّم هذا البحث أسلوباً يستند إلى نظريّة المخطّطات لتلخيص النّص العربي القائم على نموذج الرّسم البياني وخوارزميّة التّرتيب أو التّصنيف. واعتمد هذا البحث أولاً على إنشاء الرسم البياني لتمثيل بنية المستندات أو الوثائق النّصيّة، ثم قمنا بإجراء تصنيفٍ على امتداد هذا الشّكل البياني. ولغرض تحسين أداء الأسلوب المقترح، قُمنا باستخدام طريقة «الحدّ الأقصى الهامشي للارتباط "Maximal Marginal Relevance" للتخلُص من التّكرار، كما قُمنا بمُقارنة أسلوبنا المقترح مع الأنظمة الأخرى المطبّقة على الوثائق العربيّة. وتُشير جميع نتائج التّجارب التي أجريناها على هذا النّظام إلى أن أسلوبنا المقترح يمكن تطبيقه - بشكل جيّد - على اللُّغة العربيّة دون الحاجة لمعرفة مُعمّقة بتقنيّات مُعالجة اللُّغة الطبيعيّة. بالإضافة إلى ذلك، ليست هناك حاجة إلى بيانات للتّدريب. وقد أظهرت نتائج تقييم النظام باستخدام مقياس اختبار-F، أنّ الأسلوب المقترح يتفوق على الأنظمة الأخرى الموجودة.

وأوصت الدراسة أن يُتعامل في المستقبل مع مسألة كيفية تحسين الأداء في سياق تلخيص النص العربي من خلال توخي مزيد من تطوير هذا العمل في عدة اتجاهات. ويتمثّل أحد هذه الاتجاهات في اختبار هذا الأسلوب على ذخائر لُغويّة مُوسّعة تتضمن المزيد من الوثائق.

World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 3, 62 – 67

Year of publication: 2012

Arabic Text Summarization Model Using Clustering Techniques

Ahmad Haboush, Maryam Al-Zoubi, Ahmad Momani, Motassem Tarazi

نموذج لتلخيص النّص العربي باستخدام تقنيات العنقدة

مقدمة

يأتي هذا العمل في سياق استقصاء وضع نموذج لتلخيص النّص العربي التلقائي، وقد استُخدمت في هذا النموذج تقنية التّجميع العُنقودي لجذور الكلمات كنشاط رئيسي. وخلافاً للأنظمة التي طُوّرت سابقاً لتلخيص النّص العربي، والقائمة على أسلوب الاستخراج، يعتمد النموذج الحالي على وزن عُنقود (cluster weight) جُذور الكلمات العربيّة بدلاً من وزن الكلمة نفسها.

ووضّحت هذه الورقة النموذج بدقة من خلال مراحله المختلفة، فمن الواضح أنّ المخطط العام للنموذج يتبع أسلوب النموذج الوصفي التقليدي لمعظم مراحل النظام في الأدبيات ذات الصّلة، باستثناء مرحلة الترتيب أو التّصنيف. وقد أُخضِع هذا النموذج وتقنيته المبتكرة لمجموعة من التّجارب، واستُخدمت أمثلة مُختلفة للنّص العربي لأغراض التقييم، كما احتُسبت كفاءة التلخيص من حيث مقاييس الدّقة والاستدعاء. وفي الواقع تُعدُّ النتائج التي تمّ الحُصول عليها واعدة ومُنافسة لطريقة «تصنيف الفعل/ الاسم»، وما يُعزّز هذه الحقيقةَ النتائجُ التي أظهرتها الدّراسة، حيث بلغت نسبة الدّقة ٧٦٪ والاستدعاء ٧٩٪ في مُقابل نظيرتها طريقة تصنيف الاسم/ الفعل، حيث حصلت على نسبة ٦٢٪ و ٧٠٪ لكُلٍ من الدّقة والاستدعاء على التوالي. ويُعزى الفضل في هذه النتيجة الملموسة إلى الإدماج الضِمني (implicit embedding)

للقُدرة الدلالية (semantic capability) في النموذج المطوّر، لتوسيع حدود المادة المستخلصة إلى أقصى نهايات مُمكنة لِفكرة المخطّط.

المُبررات ومُشكلة البحث

تسمح طبيعة اللُّغة العربيّة والمجموعة الواسعة من مُشتقات الكلمة الوظيفية باستخدام مستوى أعلى من الاستقصاء النّحْوي. وبالتالي، يمكن إنشاء جُمل مفاهيميّة (conceptual sentences) مماثلة، إما عن طريق الكلمات النّظيرة (analogous) أو الكلمات غير المتشابهة لصِياغة التعبير (expression formalization)، وهو ما يسمح بإعطاء فُسحة لنطاق أوسع من الاستقصاء لاعتماد المخطط التجريبي المعتمد على المادة المستخلصة والتلخيصيّة بشكل متزامن. وقد استُقصيت هذه الحقيقة من خلال هذا العمل لاقتراح نموذج للتلخيص التلقائي للنص العربي يعتمد على مستوى مُنخفض من فكرة التلخيص الموجّهة من خلال مُخطط الاستخراج.

النتائج والتوصيات

تُقدّم هذه الورقة نموذجاً جديداً للتّلخيص التلقائي للنّص العربي، والميزة الرئيسية لهذا النموذج هي القدرة على تأصيل الكلمة (أو الإتيان بجذرها)، وهو ما يجعل هذا النموذج أقرب إلى الأسس الدلالية منه إلى النّحوية.

وفي الواقع تعدّ النتائج التي حُصِّلت واعدة ومُنافسة لطريقة تصنيف الفعل/ الاسم. وتعزّز نتائج هذه الدّراسة ذلك، حيث بلغت نسبة الدّقة ٧٦٪ والاستدعاء ٧٩٪ في مُقابل نظيرتها «طريقة تصنيف الاسم/ الفعل»، حيث حصلت على نسبة ٦٢٪ و ٧٠٪ لكُلٍ من الدّقة والاستدعاء على التوالي. ويُعزى الفضل في هذه النتيجة الملموسة إلى الإدماج الضمني للقُدرة الدلالية في النموذج المطوّر لتوسيع حدود المادة المستخلصة إلى أقصى نهايات مُمكنة لِفكرة المخطّط.

International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, 2014, Putrajaya, Malaysia

Year of publication: 2014

Automatic Arabic Text Summarization Using Clustering and Keyphrase Extraction

Hamzah Noori Fejer, Nazlia Omar

تلخيص النص العربي التلقائي باستخدام العنقدة واستخراج العبارة الرئيسية

يُعرف تلخيص النصوص بأنه شكل مبسّط وموجز للنص الأصلي يبين أهم الأفكار فيه. ويحتاج التلخيص للفهم المسبق للنص الأصلي ليكون شاملاً وموضّحاً للأفكار المهمة في النص، لكن ليس من السهل استخدام الحاسوب لمثل هذه المهمة، فمن الصعب جدا على الحاسوب أن يستطيع فهم نص كامل وتلخيصه، لذلك تمتاز غالباً الجمل الرئيسية من النص الأساسي وتقدم على شكل ملخص للنص، وتعدّ هذه الطريقة من أكثر الطرق شيوعاً في تلخيص النص بشكل آلي (Automatic Text Summarization).

ويعتمد البحث الآلي عموماً على طريقتين أساسيتين: الطريقة الأولى تسمى التلخيص الاستخراجي (Extractive summarization)، وتكون هذه الطريقة محدودة جداً بحيث إنها فقط تستخلص الجمل المهمة وترتبها بتسلسل زمني لضمان التماسك بين أجزاء النص. الطريقة الثانية تسمى التلخيص التجريدي (Abstractive summarization)، وهذه الطريقة تلخص النص بشكل أوضح وأكثر فهماً، فلا تعتمد على حرفيّة اللغة في تلخيص النص الأصلي، حيث يُستغنى عن بعض الكلمات الغامضة وغير المفهومة ويستبدل بها كلمات أكثر وضوحاً. وعلى الرغم من النتائج المرضية التي توصل لها الباحثون في هذه الطريقة، التي تغلبت على الطريقة الأولى من حيث الكفاءة إلا أنها طريقة معقّدة جدا من حيث التطبيق، لذلك اعتمد الباحثون على الطريقة الأولى في التلخيص نظرا لسهولة التعامل معها.

يهدف الباحثون في هذا البحث إلى إيجاد طريقة جديدة للقيام بهذه المهمة بشكل آلي وفعال، فقاموا باعتماد طريقة تسمى طريقة التجميع الهجين (hybrid clustering method) ، وتقوم على تجميع بعض الوثائق العربيّة وتقسيمها إلى عدة مجموعات، ومن ثم استخراج العبارات الرئيسية والمتشابهة من كل مجموعة على حدة من خلال تطبيق نموذج لاستخراج تلك الجمل سُمّي «وحدة استخراج العبارة الرئيسية» (key phrase extraction module)، ويقوم هذا النموذج باختيار واحدة من تلك الجمل المتشابهة وحذف الجمل الباقية استناداً إلى استخدام عدة خوارزميات لقياس درجة التشابه (مثل تشابه جيب التمام (cosine similarity) ومعامل جاكارد (Jaccard coefficient)) واختيار الأفضل منها اعتماداً على النتيجة. ولأغراض التقييم اعتمدت الطريقة المقترحة على مصفوفة (Understudy for Recall Oriented Gisting Evaluation (ROGUE)). كما استُخدمت الذخيرة اللغوية للملخصات اسيكس العربيّة (Arabic Summaries Corpus Essex) لأغراض فحص فعالية الطريقة المقترحة، حيث تحتوي هذه المجموعة على العديد من المقالات العربيّة في مواضيع مختلفة سبق تلخيصها بشكل يدوي.

تميّزت الطريقة المقترحة بأنها يمكن استخدامها على نطاق مستند واحد أو على عدة مستندات معاً. وقد قُورنت بغيرها من الطرق السابقة، وحققت فعالية عالية فيما يتعلّق بمستوى الدقة (Accuracy) على نطاق تلخيص المستند الواحد، أي ما يقارب (٨٠٪) ،أما على نطاق تلخيص المستندات المتعددة فقد حققت مستوى جيدا من الدقة وصل إلى(٦٢٪).

8th International Conference on Informatics and Systems (INFOS), Cairo, 2012, pp. NLP-7-NLP-14.

Year of publication: 2012

(Keyphrase Based Arabic Summarizer (KPAS

Tarek El-Shishtawy, Fatma El-Ghannam

مُلَخّص (برنامج تلخيص) قائم على استخراج العِبارات الرئيسيّة

المُقدّمة

مدار بحث هذه الورقة هو وصف خوارزميّة تلخيص حسابيّة عامة، للنُصوص العربيّة، رخيصَة وكفُؤة. وتنتمي هذه الخوارزميّة إلى عائلة التّلخيص الاستخراجي أو الاستخلاصي، مما يحدّ من ظهور مُشكلة التعرف على الجُمل النموذجيّة واستخراج المشاكل الفرعيّة. تُحدد العبارات الرئيسية للوثيقة التي ستُلخّص باستخدام توليفات من السمات الإحصائيّة واللُّغويّة. وتستفيد الخوارزمية مدار بحثنا من العبارات الرّئيسية باعتبارها السّمات أو الخصائص الأساسية لتصنيف الجُملة. ويوضّح هذا العمل التّجريبي تقنيات مختلفة لتحقيق أهداف التلخيص المختلفة بما في ذلك: ثراء المعلومات، وتغطية كل من الموضوعات الرئيسية والفرعيّة (المساندة)، وإبقاء التكرار بالحدّ الأدنى الممكن. بعد ذلك جرى تبنّي مُخطط منح النقاط أو التّصنيف الذي يوازن بين أهداف التّلخيص هذه. ولغرض تقييم الملخّصات العربيّة النّاتجة بالمقارنة مع أنظمة راسخة، استُخدمت النُصوص الإنجليزية / العربيّة المتحاذية خلال التّجارب.

مُشكلة البحث

تستعرض هذه الورقة تقنية تلخيص حاسوبية عامة كفُؤة، ترتكز إلى مبدأ التلخيص الاستخراجي القائم على العبارة الرئيسية. وفي حين تُستخرج العبارات الرئيسية تلقائياً من

النّص الموجود في الوثيقة، تُستخدم هذه العبارات المستخرجة لتقييم أهمية كل جُملة في الوثيقة. وعلى الرغم من وجود العديد من تقنيات الاستخراج على مستوى الجُملة، إلاّ أنه لا يولى إلاّ القليل من الاهتمام لتغيير استراتيجية الاستخراج لتحقيق واحد أو أكثر من أهداف التلخيص. فالملخِّص البشري لديه القدرة على تحديد الجمل التي ستُقدّم وفقاً لعوامل كثيرة، بما في ذلك الحد الأقصى المسموح به من عدد الجُمل التي ستُعرض وعدد الموضوعات التي ستُغطى في الوثيقة. كما أنّ الملخّص البشري يمكنه تغيير استراتيجية الاختيار إذا لاحظ أنّ الوثيقة تحتوي على مفاهيم رئيسية أو مهمة متساوية. لذا، فالهدف الرئيسي من هذا العمل، الذي يُعد الأوّل من نوعه، هو توضيح الحدّ الأدنى من التراكيب اللغُوية (العبارات) واستخدامه، وهي طريقة أكثر مرونة في توجيه مُستخرِج الجُمل المقترح نحو تحقيق واحد أو أكثر من الأهداف التالية:

- استخراج الجُمل الأكثر ثراءً بالمعلومات التي تتضمن الموضوعات الرئيسية.
- استبعاد سيطرة موضوع واحد على الملخّص الناتج.
- إبقاء تكرارات الجُملة ضمن الحدّ الأدنى.
- تغطية جميع الموضوعات المهمة الموجودة في الوثيقة.
- تحقيق التوازن بين جميع الأهداف السابقة.

الاستنتاج

في هذا البحث قدمنا خوارزميّة تلخيص عربيّة لاستخراج الجُمل ذات الصِلة من النُصوص الحُرّة. ويعتمد هذا النظام على استغلال السِمات الإحصائية واللُّغوية لتحديد العبارات الرئيسية، ومن خلال التجارب التي أجريناها، بينّا كيف أن مُخطّطات منح النقاط أو التّصنيف المختلفة القائمة على أساس العبارات الرئيسية يمكن أن توجه مُستخرج الجُملة المقترح نحو واحد أو أكثر من أهداف التلخيص. ويمكن لهذه الأهداف أن تكون غنية بالمعلومات، وأن تغطّي الموضوعات الرئيسية والفرعيّة، وأن تحدّ من التكرار. واعتمد هذا العمل مُخطط تصنيف يوازن بين أهداف التلخيص المختلفة.

Computer Speech and Language 26 (2012) 260–273

Year of publication: 2012

Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization

Houda Oufaida, Omar Nouali, Philippe Blache

التكرار الأقل والارتباط الأكبر لتلخيص ملفات النصوص العربيّة المفردة والمتعددة

إنّ عملية التكرار الأقل والارتباط الأكبر (Minimum redundancy maximum relevance (mRMR)) هي عملية تحليل تمييزي تهدف لاختيار مجموعة جزئية من الخصائص تمثل الخصائص كاملة على أفضل نحو، وتعتمد على معلومات مشتركة بين أزواج الخصائص تعكس مستوى التشابه بينها، فإذا كان هناك مقدار كبير من المعلومات المشتركة بين خاصيتين فإننا نصفها بأنّها مترابطة بشكل كبير، وبالتالي يمكن استبدال إحداها بالأخرى بأقلّ قدر من فقدان المعلومات.

تُستخدم المعلومات المشتركة لقياس التكرار والارتباط، ولتقليل التكرار فإننا نهتم بإيجاد الخصائص غير المتشابهة، أما زيادة الارتباط فيتطلب اختيار الخصائص التي تمثل مجموعة البيانات بأفضل شكل، وذلك في خطوتين هما: تطبيق طريقة تصنيف لإيجاد فئات مختلفة من الملاحظات، ثم العثور على المعلومات المشتركة بين الخصائص ومتغيرات التصنيف الناتجة من إحداهما.

تلخيص نص عربي واحد

إنّ تكرار المصطلحات أمر يتعلق بالنص الأصلي وهو عائق يريد كلّ نظام تجنبه في الملخصات الناتجة، لذا نقترح استخدام الطريقة المشروحة سابقا. إنّ قياس الارتباط لجملة ما

هو الأمر الجديد الرئيس في طريقتنا، وهي تعتمد على ارتباط المصطلحات في الجملة نفسها، ولتحقيق ذلك نستخدم أولا خوارزمية تجميع لتجميع الجمل المتشابهة في مجموعات.

تحضير النص

الجملة في نظامنا هي وحدة الاستخلاص والمصطلح هو وحدة التسجيل، لذا لا بدّ من تحضير النص الأصلي بالقيام بعدة خطوات هي: تقسيم الجمل، والتقطيع (Tokenization) وحذف كلمات التوقف، واستخلاص الجذور.

تسجيل المصطلحات باستخدام طريقة التكرار الأقل والارتباط الأكبر

في هذه الطريقة تُسجّل الخصائص على أساس كمية المعلومات المميزة التي تحملها، وفي التلخيص نهتم بالمصطلحات ذات التميّز العالي، التي تسمح لنا باختيار جمل محددة وليس غيرها، وإيجاد هذه المجموعة من المصطلحات هو ما سيحققه لنا استخدام طريقة التكرار الأقل والارتباط الأكبر.

خوارزمية استخلاص الجمل

آخر عملية يجريها نظام التلخيص هذا هي استخلاص أعلى الجمل بناء على حجم التلخيص المطلوب (نسبة الكبس وعدد الجمل وعدد الحروف)، ونقترح هنا خوارزمية استخلاص تأخذ بعين الاعتبار المصطلحات في الجمل المختارة أصلا لاحتساب نتيجة الجملة التالية لضمها إلى الملخص.

التكرار الأقل والارتباط الأكبر لتلخيص عدة نصوص

يكون تقييم الجمل هنا أكثر صعوبة بسبب الاحتمالية العالية لتداخل المعلومات عبر الجمل (cross sentence informational subsumption) إذا كان الملف يناقش نفس الموضوع، وذلك يعكس حقيقة أن بعض الجمل تعيد نفس المعلومات الموجودة في جمل أخرى. ولأننا

نستخدم المصطلحات عنصرَ القياس الرئيس، فإنّ طريقة التكرار الأقل والارتباط الأكبر يمكن استخدامها بسهولة لتلخيص عدة نصوص، ولا يدخل موقع الجمل في معادلة القياس، ويمكننا بالتالي استخدام إحدى طرائق التجميع: كلّ نص مقابل لمجموعة، أو استخدام «نموذج حقيبة الجمل» (model sentences-of-bag) وتطبيق خوارزمية تجميع، لإنتاج مجموعات وتحديد معامل التصنيف، أما الخطوات التالية فتكون كما هي في تلخيص نص واحد.

بعد اختبار طريقتنا في تلخيص نص واحد وعدة نصوص كانت النتائج الكلية واعدة، وقد تفوّقت هذه الطريقة في تلخيص النص الواحد، أما في تلخيص عدة نصوص فقد قارنّاها بنظام آخر وكانت النتائج مقبولة.

International Conference on Asian Language Processing (IALP), Tainan, 2016, pp. 26-29. doi: 10.1109/IALP.2016.7875927

Year of publication: 2016

Towards Building a Standard Dataset for Arabic Keyphrase Extraction Evaluation

Muhammad Helmy, Marco Basaldella, Eddy Maddalena, Stefano Mizzaro and Gianluca Demartinit

نحو بناء قائمة بيانات قياسية لتقييم استخلاص العبارة الرئيسيّة في النّص العربي

المُقدّمة

العبارات الرّئيسية هي عبارات قصيرة تمثّل محتوى المستند أو الوثيقة على أفضل وجه. ويمكن أن تكون مفيدة لدى طيف واسع من التطبيقات، بما في ذلك نماذج تلخيص الوثائق واسترجاعها. في هذه الورقة، نقدّم قائمة البيانات الأولى من العبارات الرئيسيّة لمجموعة من الوثائق العربيّة، التي يتمّ الحُصول عليها عن طريق أسلوب حشد الموارد "crowdsourcing". وقد قمنا من خلال التجربة بتقييم استراتيجيات تجميع الردود من خلال أسلوب حشد الموارد ، والتحقق من أدائها في مُقابل شرُوح الخُبراء لتقييم جودة قائمة البيانات الموجودة لدينا. واستعرضت هذه الورقة نتائجنا التّجريبية، وسمات قائمة البيانات، وبعض الدروس المستفادة، والأفكار التي يمكن العمل عليها في المستقبل.

المُبررات ومُشكلة البحث

مع أنّ مشكلة الفهرسة التلقائية للوثائق باستخدام الكلمات الرئيسية قد دُرِست على مدى أكثر من خمسين عاماً، غير أن مُهمة الاستخراج التلقائي للعبارة الرئيسية لفتت انتباه المجتمع البحثي فقط في أواخر التّسعينيات. ومنذ ذلك الحين، طُوّرت العديد من خوارزميات الاستخراج

التلقائي للعبارة الرئيسية في الجُملة، ويمكن تصنيفها إلى فئتين: الأسلوب الخاضع للإشراف أو الرقابة (supervised)؛ والأسلوب غير الخاضع للإشراف (unsupervised). ويتطلب الأسلوب الخاضع للإشراف وجود قائمة بيانات تدريبية لخوارزمية التعلم الآلي الخاصة به، في حين يتطلب كلاهما معيارا ممتازا لتقييم العبارات المستخرجة. وقد اقترحت العديد من قوائم البيانات خلال السنوات الماضية، وبصورة بارزة تلك الخاصة باللُّغة الإنجليزية حصراً. وعلى حدّ علمنا، فإنّ جميع قوائم البيانات المنتجة بشكل جيد والمتاحة تغطي فقط اللُّغة الإنجليزية. ممّا أدى بوضوح إلى تأخير تطوير مجموعة استخراج العبارة الرئيسية متعددة اللُّغات.

أسلوب البحث

استخدمنا في هذا البحث منصة حشد الموارد "Crowdflower" لتكوين مجموعة استخراج عربيّة للعبارة الرئيسية، وذلك بدعم من ٢٢٦ عاملاً.

الاستنتاجات والتوصيات

في هذه الورقة، استعرضنا جهودنا الأولى في بناء قائمة بيانات جديدة للوثائق العربيّة اعتماداً على أسلوب "حشد الموارد ". وعلى اعتبار أنّ هذا العمل هو الأوّل من نوعه في بناء مثل هذه الذخيرة اللُّغويّة، هناك الكثير من الاتجاهات لاستكشافها في المستقبل. وهناك إمكانية لتوسعة حجم هذه الذخيرة في المستقبل بإضافة المزيد من الوثائق؛ لكن قبل القيام بذلك، علينا دراسة بعض المسائل بمزيد من التفصيل. على سبيل المثال، نعتزم تجربة أساليب ومتغيرات مُختلفة لتَصفية العبارات الرئيسية عالية الجودة. كما أنّه من المهم أيضاً فهم العدد المثالي من العمال المطلوب لكل مُستند أو وثيقة؛ ففي هذه التجربة استخدمنا ١٠ عُمال. وقد يكون أول اتجاه للبحث يدور حول معرفة ما إذا كان هناك بعض التقنيات لأخذ العيّنات التي يمكن أن تؤدي إلى استخلاص العبارات الرئيسية بدقّة، مع استخدام أعداد أقلّ من العاملين، وبالتالي، تقليل التكلفة. وأخيراً، وفي ذات الإطار هناك خُطط مُستقبلية أيضاً لمحاولة التعامُل مع تصاميم تجريبية مُختلفة.

٣-٤-٩ أبحاث التحليل الدلالي

وتضم سبعة أبحاث بينها بحث واحد نوع (أ) هو: الإسهام في التحليل الدلالي للغة العربيّة، وستة أبحاث نوع ب، هي: استخلاص المعلومات الدلالية: دراسة مقارنة تجريبية لأدوات معالجة اللغات الطبيعية وموارد اللغة العربيّة، وطريقة لغويّة لوسم الأحكام المعيارية العربيّة على أساس الاستكشاف السياقي، و بناء آلي لمعجم مشاعر ذي نطاق واسع للغة العربيّة الفصحى والعامية، والقواعد النحوية والدلاليّة المرتبطة بإزالة اللبس عن كلمة «حتى» في اللغة العربيّة، واستخدام التجذيع والتجذيع الخفيف لقياس التشابه بين الكلمات العربيّة عن طريق نموذج التحليل الدلالي القبلي، وتصوير تشابه النصوص باستخدام ن-غرام والتحليل الدلالي القبلي.

Hindawi Publishing Corporation, Advances in Artificial Intelligence, Volume 2012, Article ID 620461, 8 pages, doi:10.1155/2012/620461

Year of publication: 2012

Contribution to Semantic Analysis of Arabic Language

Anis Zouaghi, Mounir Zrigui, Georges Antoniadis, Laroussi Merhbene

الإسهام في التحليل الدلالي للغة العربيّة

تهتم هذه الدراسة بتحديد المعنى المناسب للكلمات العربيّة الغامضة التي من الممكن أن نجدها في النصوص المقروءة من وحدة التعرف على الكلام. معظم الأساليب المتوفرة لفكّ اللبس الدلالي تختصّ باللغة الإنجليزية، فالكثير من تلك الأساليب تعتمد على موارد المعلومات التي يتم تعديلها للتمييز بين المعاني مثل:

أولا: الأساليب القائمة على المعرفة التي تعتمد على استخراج المعلومات من المعاجم وقواميس المترادفات والقواميس الآلية وشبكة الكلمات .

ثانيا: الأساليب القائمة على الذخيرة اللغوية التي تُستخدم فيها طرق إحصائية تعتمد على أعداد كبيرة جدا من النصوص، وتنقسم إلى نوعين:

التمييز القائم على الكلمة: يستخدم هذا النوع خوارزميات لقياس أوجه التشابه بعد توضيح السياقات حيث تتمثل السياقات بمساحات عالية الأبعاد تحددها الكلمات المشتركة.

التمييز القائم على الرموز: يعمل على تجميع السياقات التي تحتوي على الكلمة الهدف بحيث تحمل هذه الكلمة نفس المعنى في تلك السياقات.

في هذه الدراسة اقترحت بعض الخطوات لبناء نظام فك اللبس الدلالي للغة العربيّة، حيث استخدمت نسخة إلكترونية من المعجم الوسيط لبناء قاعدة بيانات تحتوي على الكلمات ومعانيها. وبعد ذلك حذفت كلمات التوقف من الجملة الأصلية (التي لا يؤثر حذفها على

المعنى) باستخدام قائمة الحذف (stopwords) المحددة مسبقا في قاعدة البيانات المستخدمة. كما استُخدمت الكلمات المكوّنة للجمل الموظّفة لتعريف الكلمة الغامضة في المعجم لاستخراج السياقات المختلفة لتلك الكلمة من الذخيرة اللغويّة، وذلك بجمع نتاج عمل خوارزميات استخراج جذور الكلمات بخوارزميات مطابقة السلسلة التقريبية لإيجاد أيّ ظهور لجذر الكلمة، ثم تخزين الناتج في قاعدة المعلومات، وعندها تمثل الجمل المحتوية على الجذر للكلمة الغامضة سياقات لتلك الكلمة.

وقد استُخدمت خوارزميات الشلبي وكنعان لاستخراج جذور الكلمات العربيّة؛ لأنّها لا تستخدم أيّ مصادرَ بل تقوم بتعيين وزنٍ لأحرف الكلمة، وتعتمد تلك الأوزان على أرقام حقيقية محددة من صفر إلى خمسة مضروبة بقيمة معيّنة تعتمد على موقع الحرف في الكلمة.

أما الخطوة المقترحة التالية فهي قياس نسبة التشابه بين السياقات المختلفة التي استُخرجت من تعاريف الكلمات الغامضة في المعجم والسياقات الحالية المستخرجة من الذخيرة اللغويّة، فالسياق الحاصل على أعلى نسبة تشابه مع السياق الحالي يعطي المعنى الأكثر ملائمة لتلك الكلمة الغامضة.

وفي النهاية قمنا بتطبيق الأنظمة المستخدمة في مجال استرجاع المعلومات التي تقوم بمقارنة الجملة الأصلية أو السياق الحالي بالسياقات المختلفة التي استُخرجت للكلمة الغامضة. كما قمنا بتعديل خوارزميات مبسّطة من خوارزميات ليسك، لتعيين المعنى الصحيح من بين المعاني المختلفة المقترحة من تلك الأنظمة.

النتائج التجريبية:

أجريت اختبارات لمعرفة مدى صلاحية الخوارزميات المستخدمة بالاستعانة ببعض الأدوات المتاحة مجانا. ففي دراسات مشابهة خاصة باللغة الإنجليزية عادة ما يُقيّم العمل باستخدام Senseval-1 أو Senseval-2، بينما في دراستنا الحالية كان يجب تجهيز بياناتنا التجريبية الخاصة باستخدام مجموعة من الموارد المختلفة تماما.

لقياس معدل الغموض استخدمنا أكثر تقنيات التقييم شيوعا، وتعمل على اختيار عيّنة صغيرة من الكلمات ومقارنة نتائج النظام مع نتائج الإنسان. كما استخدمنا مقياسا للدقة ومقياسا للاسترجاع ومقياس F-score المتوازن الذي يحدد المتوسط التوافقي المرجّح للدقة. وكحد أعلى قمنا باستخدام السياق الذي يتوافق مع المعنى الأكثر شيوعا.

الأدوات المستخدمة والبيانات التجريبية:

معجم: استُخدمت نسخة إلكترونية من المعجم الوسيط لبناء قاعدة بيانات تحتوي على الكلمات ومعانيها، حيث اخترنا العمل على المعاني بالغة الدقة مما يجعل عملنا صعبا ومعقّدا لأنّه يزيد من عدد المعاني التي سوف نأخذها بعين الاعتبار.

ذخيرة لغوية: اخترنا العمل على نصوص من مختلف المجالات مثل الرياضة، و السياسة، والدين، والعلوم، وغيرها. واستُخرت تلك النصوص من مقالات في الصحف المسجّلة في ذخيرة السليطي-أتويل. تحتوي الذخيرة اللغويّة المستخدمة على ١٥٠٠ نص بمعدل ٥٠٠ كلمة/ النص. كما تحتوي الذخيرة على ٥٠ كلمة غامضة، وكان متوسط عدد المترادفات لكل كلمة غامضة هو ٤، أما متوسط عدد المعاني المحتملة فهو ١٢. كما أنّ متوسط حجم كل سياق استخدام هو ٩٧٠ كلمة و ١٣٠ جملة.

تتميز هذه الوثائق بامتلاك هيكل واضح ليسهل عرضها واستخدامها في سياقات مختلفة للعثور بكفاءة عالية على كلمات ذات صلة.

قائمة الحذف (Stopwords): جُمعت قائمة بالكلمات غير المرغوب بها في عملية البحث التي لا يؤثر حذفها على معنى الجملة مثل بعض الضمائر والأسماء والأحرف والمصادر وبعض الكلمات التي تُعدُّ غير مهمة واحتوت تلك القائمة ٢٠٠٠ كلمة.

البيانات التجريبية: اختيرت ٥٠ كلمة غامضة وقُيِّم ٢٠ مثالا لكل معنى يمكن أن تحمله الكلمة الغامضة. ويمكن أن يرى بعض الدراسين أنّ ٥٠ هو عدد غير كاف بسبب التحديات التي واجهناها خلال تلك التجربة، مثل: عدد المعاني المعطاة في المعجم لتلك الكلمات، صعوبة

تجزئة بعض الجمل العربيّة بسبب اللبس فيها، بالإضافة إلى أهمية إيجاد عيّنات اختبار يمكن الحكم عليها وتكون مختلفة بدرجة غير كبيرة.

النتائج التي حصلنا عليها:

١. أثر استخراج الجذور ومطابقة السلسلة : قمنا بقياس أداء نظامنا باستخدام المقياس المقترح سابقا مع وبدون استخدام خوارزميات استخراج الجذر وخوارزميات مطابقة السلسلة على التوالي.

٢. أثر حجم السياقات المستخدمة: لتحديد حجم سياقات المعاني المختلفة للكلمة الغامضة قمنا بتقييم النتائج المعطاة من نظامنا المقترح وذلك مع تغيير حجم السياق بين ٥٠ كلمة، و ١٠٠ كلمة، و ١٥٠ كلمة، فاستنتجنا أنّ أقل معدل لفكّ اللبس عائد بالدرجة الأولى إلى عدد السياقات غير الكافي الذي يؤدي إلى عدم تلبية جميع الأحداث الممكنة، لذلك حاولنا جمع أكبر عدد ممكن من النصوص لتوسيع حجم قاعدة المعرفة خاصتنا.

٣. أثر حجم نافذة السياق: في دراسة لأثر حجم نافذة السياق على عملية فكّ اللبس الدلالي للكلمة تبين أنّ الكلمات الرئيسية الموجودة ضمن السياق الأصغر والمكوّن من ست إلى ثماني كلمات هي أكثر الكلمات إفادة لفك اللبس الدلالي، و أشرنا إلى صعوبة تمييز العناصر الرئيسية التي تحدد معنى الكلمة في السياق الطويل جداً.

لقد حددنا حجما مثاليّا للسياق المناسب لكل اختبار وذلك لحل مشكلة عدم وجود حجم ثابت لنافذة السياق لكل الكلمات. فعلى سبيل المثال قمنا باستخدام سياق بحجم ثلاث كلمات (أي ثلاث كلمات إلى يسار الكلمة الغامضة، وثلاث كلمات إلى يمينها)، وسياق آخر بحجم كلمتين وآخر بحجم كلمة حيث سُجّلت أفضل نتيجة لقياس تشابه السياقات عند استخدام السياق بحجم ثلاث كلمات.

الخلاصة:

حققت الخوارزميات المقترحة ما نسبته ٧٨٪ في الدقة ونسبة ٦٥٪ في الاسترجاع. ونقترح أن يتم مستقبلا تقليل نسبة التطابق بين الكلمات ومعانيها لبناء نظام قائم على قواعد خاصة لفك لبس الكلمات العربيّة.

IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, 2016, pp. 879-887

Year of publication: 2016

Semantic Information Retrieval: A comparative experimental study of NLP Tools and Language Resources for Arabic

Nadia Soudani, Ibrahim Bounhasa, Yahya Slimania

استخلاص المعلومات الدلالية: دراسة مقارنة تجريبية لأدوات معالجة اللغات الطبيعية وموارد اللغة العربيّة

مع التطور الكبير الذي ظهر في السنوات الأخيرة في مجال استخدام الإنترنت، وزيادة عدد المستخدمين وبالتالي ازدياد المعلومات المتوفرة على الشبكة، كان لابد من الاهتمام بكلمات البحث للحصول على ما يريده المستخدم بشكل دقيق وواضح. وبالنسبة لبعض اللغات فقد تمّ التغلب على هذه المشكلة، ولكنّ الأمر ما زال قيد الدراسة بالنسبة للغة العربيّة، لكونها لغة معقدة من ناحية نحوية، ودلالية وصرفية.

هدف الباحثون في هذه الدراسة إلى خدمة اللغة العربيّة، التي تتصف بأنها لغة غنية بالدلالات، وأفضل خدمة هنا هي في عملية استرجاع المعلومات (**Information Retrieval**) ، حيث تستخرج هذه الدلالات من القواميس أو المعاجم العربيّة، ومن ثمّ يتم تحليلها والتفكير فيها عن طريق استخدام معالجة اللغة الطبيعيّة (**Natural Language Processing**) وأدوات استخراج النصوص (**text mining tools**)، مما يساعد في فهم السياق وتحديد معنى كلمات الاستعلام (**query terms**) المستخدمة في عمليات التصفح عبر الإنترنت، للحصول على نتائج ذات صلة عند البحث. اعتمد الباحثون على توسيع الاستعلام الدلالي(**Semantic Query Expansion**) للمساعدة في تطوير عمليات البحث وتقليص الفجوة بين ما تعنيه كلمة الاستعلام وما يبحث عنه المستخدم. وكذلك قام الباحثون بدراسة تأثير علم الصرف في اللغة

العربيّة على التفسير الدلالي للكلمات في عمليات الاستعلام.

في هذه الورقة البحثيّة، اهتم الباحثون بالبحث الدلالي في محتوى اللغة العربيّة، فقدموا نهج البحث الدلالي العام القائم على استخدام دلالات الكلمات، وقد أجرى الباحثون دراسة تجريبية لمقارنة أدوات معالجة اللغة الطبيعية فيما يتعلق باللغة العربيّة واستخدامها في الموارد اللغويّة التي أظهرت تأثيرا واضحا في أداء البحث الدلالي في محركات البحث المختلفة. في حين يمثّل هذا العمل محاولة للتعامل مع البحث الدلالي على شبكة الإنترنت في اللغة العربيّة، وهذا يشكّل خطوة جادة لتعزيز هذه الفكرة ودراسة المقاييس التي يمكن أن تطوّر أداء البحث في الإنترنت بناء على استخدام كلمات البحث المناسبة التي تعطي دلالات واضحة. وقد استُخدمت نهج مختلفة في استخلاص المعلومات الدلالية اختبرت بناء على أدوات تحليل مورفولوجية (Morphological Analyzers) مختلفة ومصادر متنوعة من الموارد اللغويّة.

كما اختُبر النهج المقترح استنادا إلى مقاييس مختلفة، فقد اختُبر هذا النهج بناء على التحليل الدلالي وقد عرضت معدلات دقيقة وجيدة عند القيام بالبحث دون الحاجة إلى إعادة صياغة الاستعلام (Query Reformulation). ومن الواضح أن البحث القائم على الدلالات يمكن أن يحسّن الدقة ويعطي نتائج أفضل بالمقارنة مع البحث القائم على المصطلحات الرئيسيّة.

7th International Conference on Information and Communication Systems (ICICS), pp. 347-352, IEEE.

Year of publication: 2016

A linguistic Method for Arabic Normative Provisions' Annotation based on Contextual Exploration

Ines Berrazega, Rim Faiz, Ghassan Mourad, Asma Bouhafs

طريقة لغويّة لوسم الأحكام المعياريّة العربيّة على أساس الاستكشاف السياقي

قدّم هذا البحث طريقة لغويّة للتعرف التلقائي على فئات الأحكام المعياريّة (Normative Provisions) في النصوص القانونية العربيّة ووسمها بشكل دقيق. وتستند الطريقة المقترحة إلى الاستكشاف السياقي (Contextual Exploration) للتعرف على الأحكام المعياريّة المنقولة من نصوص مجهّزة ومعالجة مسبقاً. إلى جانب تصنيف الأحكام المعيارية العربيّة ومجموعة من قواعد الشروحات، فإنّ هذه الطريقة تمكّن أنظمة الوسم الدلالية (semantically annotate) من التعرّف على الأحكام المعيارية من خلال التحليل السطحي واستخراج مجموعة من العلامات اللغويّة ذات الصلة اعتماداً على السياق.

لقد اتبعت ثلاثة مفاهيم أساسية للتعرف على الأحكام المعيارية وهي؛ مفهوم العلامات اللغوية (Linguistic markers)، ومفهوم التبعيات السياقية (Contextual dependencies)، ومفهوم فضاء البحث (Search space). وفيما يلي توضيحها:

العلامات اللغوية البحثية تنتمي إلى نوعين: المؤشرات الرئيسية والقرائن التكميلية. والمؤشرات الرئيسية هي المصطلحات التي يستخدمها المشرع للتعبير عن فئة معينة من الحكم المعياري. ومع ذلك، فإنّ هذه الخطوة لا تكفي لاتخاذ قرار بشأن الفئة المعيارية للحكم الذي تحدث فيه؛ إذ إن تحديد القرائن التكميلية أمر ضروري من أجل إزالة حالات عدم التحديد بالنسبة إلى المصطلحات متعددة الأوجه أو لحل الحالات الغامضة.

التبعات السياقية: وهذه الخطوة تعبّر عن التفسير الدلالي للمؤشرات الرئيسية وتعتمد على وجود الأدلة التكميلية أو غيابها. فعلى سبيل المثال، إذا كان المؤشر الرئيسي هو مصطلح تعدد الزوجات، فإنّ تفسيره يعتمد على وجود بعض القرائن التكميلية في السياق.

فضاء البحث: وهذا يمكن أن يكون فقرة أو جملة أو جزءا من الجملة، ويعتمد على بنية النصوص. وفي هذا العمل استخدم الباحثون الجملة للتعبير عن فضاء البحث، ومن الجدير بالذكر أنهم اعتبروا الفقرات المكونة من قائمة المسافات البادئة على أنها جملة طويلة فريدة من نوعها (وبالتالي لا بد من تحديد حكم واحد فقط).

قيُّم أداء هذه الطريقة بتطوير نموذج أولي لتحديد الفئات المعيارية تلقائيا من مجموعة كبيرة من النصوص التي جُمعت من الجريدة الرسمية للجمهورية التونسية. و كان مقياس الدقة (Precision) (٩٦,٤٪)، والاستدعاء (Recall) (٩٦,٠٦٪) ومعدل "ف" (score-F) (٩٦٫٢٣٪).

First International Conference on Arabic Computational Linguistics (ACLing), Cairo, 2015, pp. 94-99.

Year of publication: 2015

Automatic Expandable Large-Scale Sentiment Lexicon of Modern Standard Arabic and Colloquial

H. S. Ibrahim, S. M. Abdou, M. Gheith

بناء آلي لمعجم مشاعر ذي نطاق واسع للغة العربيّة الفصحى والعاميّة

إنّ تحليل المشاعر والأحاسيس ((**subjectivity and sentiment analysis** (**SSA**) مهم للتعرف على مشاعر الكاتب اعتماداً على تعليقاته سواء أكانت إيجابية أم سلبية. ويتم ذلك من خلال التحليل التلقائي لأعداد كبيرة جدا من الوثائق أو التعليقات، فتساعد الأنماط النحوية والدلالية في الكشف عن تعبيرات المشاعر أو ما يسمى بقطبية المشاعر (**polarity**) التي تُستخدم بدورها لتحديد ما إذا كان النص إيجابياً (فمثلا عبارة «إنها كاميرا مدهشة!" عبارة إيجابية، وعبارة "أكره هذا الهاتف الخلوي» تدل على السلبية، وعبارة «سأشاهد هذا الفيلم قريبا» تدل على المحايدة). تعتمد كفاءة وجودة أداء أنظمة تحليل المشاعر على مطلبين أساسيين هما؛ أولا، معجم مشاعر ذي تغطية عالية، حيث تُوضع علامات ضمن المدخلات تعمل مؤشراتٍ على التوجه الدلالي (الإيجابية والسلبية ومحايدة)، وثانياً، قاعدة بيانات مدربة لتعليم مصنف المشاعر(**classifier**).

لقد أُجريت العديد من البحوث في هذا المجال خلال العقد الماضي، ولكنّ الحاجة إلى بناء المزيد من هذه البحوث وتطويرها ما تزال قائمة، وخاصة بالنسبة للغات الغنية بالأشكال الصرفية ((**morphologically-Rich language** (**MRL**) كاللغة العربيّة. هذا البحث يعرض معجما قابلا لتوسيع نطاق تغطيته أوتوماتيكياً لكلمات المشاعر العربيّة، ويهدف إلى توفير مورد معجمي مبني بشكل متخصص لدعم تصنيف المشاعر العربيّة وتطبيقات استخراج الرأي. وقد

بُني المعجم باستخدام نواة من المعاني المعيارية للكلمات العربيّة التي جُمعت وسُميّت يدويا اعتماداً على توجهها الدلالي (الإيجابي أو السلبي)، حيث يكون التوسع التلقائي للتنبؤ باتجاه المشاعر عن طريق كلمات المشاعر الجديدة التي يتم توليدها من خلال استغلال بعض الأساليب المعجمية مثل أقسام الكلام ((POS) speech-of-part) وتقنيات تجميع المترادفات (Synonyms) من المعاجم العربيّة المتوفرة مجاناً على الإنترنت، وموسوعات المفردات (thesauruses) العربيّة المختلفة.

وهكذا فإنّ الجهود المبذولة في هذا البحث سعت لتوسيع المعاجم القطبية العربيّة باستخدام أنواع مختلفة من البيانات، حيث ركّزت البيانات على التغريدات (tweets) المدوّنة باللغة العربيّة الفصحى واللهجة المصرية ومواقع الويب لبعض المدونات والمنتجات (مثل حجز الفنادق، وعروض المنتجات، وتعليقات البرامج التلفزيونية). وأظهرت نتائج التجارب أنّ طريقة التوسيع الآلية المقترحة أسفرت عن أداء عال وكفاءة في الكشف عن كلمات المشاعر وتحديد القطبية. كما تشير النتائج إلى أنّ عدد الإدخالات الكليّة في المعجم زادت تدريجيّاً، وبالتالي زادت تغطية معجم القطبيّة، ولكن زادت الفجوة في النتائج الخاصة باللهجة العاميّة حيث تراوحت الزيادة ما بين (٢٪ - ٥٪) وذلك يعود إلى تزايد كلمات اللهجة العامية المستخدمة في الحياة اليومية.

5th International Conference on Information & Communication Technology and Accessibility (ICTA), Marrakech, 2015, pp. 1-6.

Year of publication: 2015

Rules-based grammatical and semantic disambiguation of the token "ḥattā" in Arabic

Dhaou Ghoul, Amr Helmy Ibrahim, Claude Audebert

القواعد النحويّة والدلاليّة المرتبطة بإزالة اللبس عن كلمة "حتى" في اللغة العربيّة

تعرض هذه الورقة البحثية طريقةً للتوضيح النحوي والدلالي (grammatical and semantic disambiguation) لكلمة (حتى) "ḥattā" في اللغة العربيّة. وتستند الطريقة المقترحة إلى تحليل شامل للنص من خلال السياق بهدف الوصول إلى أقصى قدر من المعلومات اللغوية لهذه الكلمة باستخدام مجموعة من البيانات من أجل النمذجة للنظام. وللقيام بذلك، طوّر الباحثون أولا مجموعة تحتوي على سياقات مختلفة من رمز "حتى". وثانيا، حددوا من هذه المجموعة المعاييرَ اللغوية المختلفة لهذا الرمز المميز التي تسمح بتحديدها بشكل صحيح. وأخيراً دوّن الباحثون هذه المعلومات بشكل قواعد لغوية من أجل الكشف عنها آلياً بكل سهولة ويسر.

تُعدُّ عملية إزالة الغموض من أهمّ متطلبات المعالجة الطبيعية للغة العربيّة ، ويمكن إزالة الغموض بثلاث طرق أساسية: الأولى إحصائيّة وتستخدم لها طرقا مختلفة مثل نموذج ماركوف (Markov model) و ن-غرام (N grams) وشجرة القرار (Decision tree) كما يمكن أن تكون الطريقة لغوية (grammatical) مثل معالجة النص حسب السياق، أو قد تكون هجينة بين الطريقتين الإحصائية والدلالية. ويعد التوضيح الدلالي واللغوي للكلمات من أبرز الطرق التي تزيد من كفاءة تطبيقات معالجة اللغات الطبيعية (مثل الترجمة الآلية وأنظمة التعرف على الصوت وغيرها)، وبشكل عام للتخلص من غموض النص لا بدّ من استخلاص الميزات من

النصوص الأصلية، والطريقة التي اتبعت هنا ركزت على السياق الذي جاءت فيه كلمة «حتى» والكلمة التالية والسابقة لها.

تضمّنت المرحلة الأولى من العمل جمع البيانات التي ستُستخدم في الدراسة، حيث تضمنت ١٥٠ جملة من اللغة العربيّة الفصحى مستخرجة من الصحف (صحيفة الوطن)، كل جملة تحتوي كلمة «حتى». وفي المرحلة الثانية صُنّفت كلمة «حتى» إلى حرف جر (prepositions) أو أداة ربط (conjunctions) حسب موقعها، وبُنيت مجموعة من القواعد اعتمادا على الموقع والسياق. على سبيل المثال عندما تكون كلمة «حتى» بين فعلين (سأدرس حتى أنجح) هنا توصف بأنها أداة ربط، في حين كلمة «حتى» في الجملة (لم تعرف القصة حتى الآن) تشير إلى ضمير. وفي النهاية قدّم هذا العمل وصفا لكلمة «حتى» في المستوى النحوي والدلالي بعد طريقة الاستكشاف السياقية. وللقيام بذلك، أنشأ الباحثون مجموعة تغطي الحالات السياقيّة المختلفة لكلمة "حتى"، ثم قاموا بتحديد قائمة غير شاملة من القواعد للحد من غموضها.

Colloquium in Information Science and Technology, Fez, 2012, pp. 69-73.

Year of publication: 2012

Stemming Versus Light Stemming for Measuring the Similarity between Arabic Words with Latent Semantic Analysis Model

Hanane Froud, Abdelmonaime Lachkar, Said Alaoui Ouatik

استخدام التجذيع والتجذيع الخفيف لقياس التشابه بين الكلمات العربيّة عن طريق نموذج التحليل الدلالي القبلي

أصبح تمثيل المعلومات الدلالية الواردة في الكلمات ضرورة لبناء أيّ تطبيق من تطبيقات معالجة النصوص العربيّة مثل القواميس الآلية وبرامج التلخيص وغيرها، وذلك بهدف استنباط الارتباطات الدلالية (**semantic dependencies**) بين الكلمات التي يمكن التعبير عنها عن طريق حساب تكرار الكلمات، وحساب كثافة توزيعها في النص، ومن أجل ذلك لا بد من حساب درجة التشابه بين الكلمات (**similarity**). جاءت هذه الورقة البحثية لتدرس أثر تطبيق اثنتين من طرق المعالجة القبلية (**preprocessing**) المطبّقة على النصوص العربيّة، وهما: التجذيع (**Stemming**) والتجذيع الخفيف ((**Light Stemming**. وفي هذا السياق نذكر أنّ التقليم هو عبارة عن إزالة زوائد الكلمات وإرجاعها إلى الجذر، مثلا كلمة (مدرسون) يمكن تقليم زوائدها وإرجاعها إلى الأصل (درس) أو (مدرس)، وذلك اعتمادا على طريقة التقليم المتبعة.

في البداية جُمعت البيانات الخاصة بالدراسة ثم نُظّمت في مصفوفة على النحو التالي: يُحسب عدد المرات التي تظهر فيها كلّ كلمة في كلّ نص، وتُوضع النتائج على صورة مصفوفة بحيث ينتج تقاطعُ الأعمدة مع الصفوف خليةً تحتوي على تكرار حدوث كلّ كلمة في النص. ومن أشهر الطرق المستخدمة لحساب المعلومات الدلالية للكلمات هي خوارزمية تحليل القيمة

المفردة (Singular Value Decomposition)، وهي طريقة عامة لتحليل مصفوفة خطيّة إلى المكونات الرئيسية المستقلة. وتسمح هذه الطريقة بتحديد ترددات مجموعة من البيانات حسب انتشارها مع مراعاة التباين بين البيانات. بعد خطوة التحليل كان لا بد من تقليل حجم البيانات (Dimensionality Reduction) بحذف غير الضرورية منها، والهدف من هذه الخطوة هو توفير الوقت مع الحفاظ على دقة النموذج المقترح. بعد ذلك جاءت خطوة التجذيع للكلمات العربيّة باستخدام طريقتين هما طريقة مجذع خوجا (Khoja Stemmer) وطريقة لاركي للتجذيع الخفيف (Larkey Light Stemmer)، وقد أظهرت النتائج التي تم الحصول عليها من البحث أنّ طريقة لاركي تفوّقت من حيث الأداء على طريقة خوجا ، لأنّ طريقة خوجا تؤثر على معاني الكلمات. ويُعزى السبب في ذلك حسب وجهة نظر الباحثين إلى أنّ طريقة التجذيع الخفيف لا ترد الكلمات إلى جذورها، فهي فقط تتخلص من السوابق واللواحق للكلمات، وبالتالي لا يتأثر معنى الكلمة كما في طريقة خوجا التي ترجع الكلمة لجذرها الثلاثي.

استُخدمت أربع طرق لقياس التشابه بين الكلمات عن طريق حساب المسافات بين متجهات الكلمات، وهي طريقة المسافة الإقليدية (Euclidean Distance) وطريقة جيب التمام (Cosine Similarity) وطريقة معامل جاكارد (Jaccard Coefficient) وطريقة معامل ارتباط بيرسون (Pearson Correlation Coefficient)، وأثبتت النتائج أنّ استخدام طريقة جيب التمام لقياس المسافة بين المتجهات هي الطريقة الفضلى، في حين كانت طريقة إقليدس الأسوأ من حيث النتائج.

SAI Computing Conference (SAI), London, 2016, pp. 269-279.

Year of publication: 2016

Visualizing Document Similarity Using N-Grams and Latent Semantic Analysis

Ashraf S. Hussein

تصوير تشابه النصوص باستخدام ن-غرام والتحليل الدلالي القبلي

إنّ ازدياد عدد موارد المعلومات وكمية الوثائق أدّى إلى تزايد الحاجة إلى تصميم أدوات ذات فاعلية وتطويرها من أجل بناء تصور يساعد في حساب مقدار التشابه بين الوثائق وكشف الانتحال وذلك اعتماداً على العلاقات الخفيّة بين المستندات. هذا البحث يقترح طريقة لتحليل محتوى النصوص وإيجاد مقدار التشابه بينها، وذلك من خلال نمذجة العلاقة بين الوثائق وتمثيل «ن غرام» (n-gram) المرتبط بكل وثيقة، حيث تعد تقنية الـ (n-gram) إحدى التقنيات المشهورة لقياس التشابه بين المستندات اعتماداً على تكرار الكلمات في المحتوى النصي للمستند، فهي تحسب بناءً على إيجاد عدد المرات التي تتكرر فيها مجموعة معينة من الأحرف في كل كلمة، فمثلاً كلمة المعلمون عند تقسيمها إلى مقاطع حرفية مع (٣- غرام) تصبح (الم، لمع، معل، علم، لمو، مون) وبعد التقسيم يتم حساب التشابه بين عدد مرات تكرار المقاطع التي توجد في كلّ كلمة.

تعتمد هذه الطريقة على تحليل محتوى الوثيقة اعتماداً على بنيتها الصرفية واشتقاقاتها المعجمية، حيث قام الباحث بتطوير طريقة جديدة للتعامل مع الملفات متعددة الأحجام، لتتم معالجة النصوص المدخلة بإجراء بعض التغييرات التي من شأنها تسهيل المهمة مثل إزالة الهمزات وتحويل «التاء» المربوطة إلى «هاء»، وتسمى هذه الخطوة التطبيع «Normalization". وبعد ذلك تُصفّى النصوص عن طريق إزالة كلمات الوقف (stop-words). وهذه الكلمات تشمل أدوات الربط وجميع المصطلحات التي لا تعطي معنى محددا (مثل أحرف الجر والعطف

وغيرها)، حيث تُزال هذه الكلمات أثناء فهرسة المستندات عن طريق فحص كلّ كلمة حسب خصائصها الصرفية، فإذا كانت لا تتطابق مع هذه الخصائص فإنها تُعدُّ من الكلمات التي يجب استبعادها. بعد ذلك، يتم تحليل النصوص عن طريق تطبيق خطوة تقسيم الكلام ((POS part-of-speech) على جميع المستندات التي فُحصت لحلّ الغموض الصرفي.

بعد خطوات معالجة النصوص التي طبّقها الباحث، وتمثيل الكلمات عن طريق تقنية الـ «ن- غرام»، بُنيت مصفوفة ثنائية لحساب تكرارات الكلمة في مستند معين وعدم تكراراها في بقية المستندات التي تسمى «مصفوفة نسب ظهور المصطلح إلى عدم ظهوره في المستند» (TF-IDF). يهدف بناء هذه المصفوفة إلى إيجاد العلاقة الخفية بين الوثائق وعبارات (ن-غرام) ، ففي هذه المصفوفة تمثل الأعمدة المستندات والصفوف عبارات (ن-غرام) . وللتحقق من هذه العلاقات استُخدمت طريقة تحليل الدلالات الكامنة ((LSA) Latent Semantic Analysis) التي تعتمد على ترابط المعاني بين الكلمات أو «المترادفات (Synonyms)"، وهذه الطريقة تُعدُّ أسلوبا رياضيا لتقليص مصفوفة متعددة الأبعاد إلى مصفوفة أقلّ حجماً، وهنا طُبّق نموذج تحليل القيمة المنفردة ((SVD) Singular Value Decomposition) لإبراز العلاقات الدلالية الموجودة بين الوثائق وعبارات (ن-غرام).

لقد قُيّمت الطريقة المقترحة عن طريق تقدير التشابه بين المستندات الأصلية ومجموعة من المستندات التي أعيدت هيكلتها بواسطة بعض التغيرات النحوية. وقد تألّفت مرحلة اختبار النموذج المقترح من تسع وثائق مستخرجة من كتاب نص القانون المصري، وكان متوسط حجم الوثائق مختلف الحجم بمعدل (٢٠٠٠) كلمة.

٣-٤-١٠ أبحاث تحليل الرأي

وتضم سبعة أبحاث بينها بحث مسحي هو: تحليل الذاتيّة والمشاعر (العواطف) في اللغة العربية دراسة مسحية، و أربعة أبحاث نوع (أ) هي التّنقيب عن الرأي العَربي القائم على السِمات الدِلالية باستخدام الأنطولوجْيا، و تَحليل المَشاعِر (العَواطف) العَربية: القائِمة على المُفردات والقائِمة على المَتن، و قياس قوة الآراء باستخدام التحليل اللغوي المعتمد على القواعد للمواقف باللغة العربية و معجم عربي واسع النّطاق للتّنقيب عن العَواطِف (المَشاعِر) في الرأي العَربي

وهناك بحثان نوع (ب): تحسين إيجاد تصنيف المواضيع وقطبياتها في المراجعات العربيّة باستخدام طرق معتمدة على القاموس والتنقيب عن وجهات النظر وتحليلها في اللغة العربية.

Advanced Machine Learning Technologies and Applications (AMLTA12)

Year of publication: 2012

Subjectivity and Sentiment Analysis of Arabic: A Survey

Mohammed Korayem, David Crandall, and Muhammad Abdul-Mageed

تحليل الذاتيّة والمشاعر (العواطف) في اللغة العربيّة: دراسة مسحية

المُقدّمة

أصبحت الشابكة اليوم تعُجّ بأعداد كبيرة من المستخدمين الذين يُعدُّونها منصّةً للقراءة والكتابة على حدٍ سواء، ولم يعد المستخدمون مُجرّد مُستهلكين للمعلومة، وإنما أيضاً مُنتجين لها. وبات محتوى ما يُنشِئُه المستخدمون في هيئة نصٍ حُرّ غير مُنظّم جُزءاً لا يتجزأ من الشّابكة، ويرجع ذلك أساساً إلى الزيادة الكبيرة في مواقع الشّبكات الاجتماعية ومواقع مشاركة الفيديو ومواقع الأخبار والمعلومات الأخرى المتاحة على الشّابكة، بالإضافة إلى مواقع المنتديات والمدوّنات. وبسبب هذا الانتشار الكبير لمحتوى ما ينشُره المستخدمون، اكتسبت هذه الثروة الهائلة من المحتوى المنشور على الشابكة اهتماماً كبيراً نظراً لأهميته بالنسبة للعديد من الشّركات والوكالات والمؤسّسات الحكومية. وأصبحت الذاتية و تحليل المشاعر (العواطف) تُشكّل جُزءاً من الفروع المهمة لدراسة المحتوى.

لقد حاز موضوع تحليل الذاتية والعواطف مُؤخّراً على الكثير من الاهتمام، لكن مُعظم الموارد والأنظمة التي وُجدت حتى الآن مُصمّمة للتعامل فقط مع اللّغة الإنجليزية وغيرها من اللّغات الهندية الأوروبية. وعليه فقد أصبح هناك حاجة مُتزايدة لتصميم أنظمة تتلاءم مع اللغات الأخرى، خاصةً وأن مواقع المدوّنات والمدوّنات الصّغيرة المنتشرة على الشّابكة باتت تكتسب شعبية في جميع أنحاء العالم. وتستعرض هذه الورقة تقنيات مُختلفة لتحليل

الذاتيّة والعواطف في اللّغة العربيّة. وبعد تقديم موجز مختصر عن اللّغة العربيّة، قدمت هذه الدراسة المسحيّة وصفاً للتقنيات الرئيسية القائمة واختبرت المتون التي وردت في الاستعراض المرجعي. وتستعرض هذه الدراسة الجُهود الرامية إلى بناء أنظمة تحليل تتعامل مع اللغة العربيّة. وتضمنت هذه الدراسة المباحث الرئيسية التالية:

أولاً: خصائص اللغة العربيّة؛ ثانياً: مراجعة للمصادر (المعاجم) واختبار لمتون اللغة العربيّة؛ ثالثاً: استعرضت الدراسة أساليب تحليل الذّاتية والعواطف في اللغة العربيّة. ثم الخاتمة.

مُشكلة البحث والمُبررات

في اللّغة الطّبيعية، تُشير الذاتية إلى جوانب اللغة المستخدمة للتّعبير عن الآراء والتقييم والتّخمين أو الاعتقادات إلى جانب المشاعر أو العواطف. وتُشير عملية التّصنيف الذاتي إلى مُهمة تصنيف النّصوص إلى نصوص موضوعية أو غير موضوعية. ويمكن أيضاً تصنيف النّص الشّخصي (غير الموضوعي) حسب معناه وجاذبيّته (استقطابه). أما عملية تصنيف العواطف أو المشاعر، فتعتمد على تحديد إذا كان النّص الشّخصي إيجابيا أو سلبي اأو مُحايدا أو مُختلطا. وقد اكتسبت هذه الأنواع المختلفة من أساليب تحليل الذات اهتماماً مُتزايداً، لأنها تُوفر طريقة تلقائية لتلخيص قدر كبير من النصوص بالنّسبة للمستخدمين (بما في ذلك المراجعين أو المدققين أو المدوّنين أو المغرّدين،...، وما إلى ذلك) في طرح الآراء التي يُعبّرون عنها. وتُعدُّ مثل هذه البيانات ذات فائدة بالنّسبة للشّركات ومُؤسّسات الأعمال التي ترغب في مراقبة مشاعر الجُمهور أو آرائه تجاه مُنتجاتها وخدماتها. ويمكن للمُواطنين أيضاً الاستفادة من هذه المعلومات في المقارنة بين المشاعر أو الآراء المختلفة حول المنتجات المتنافسة.

وبالتالي، يُمكن النظر إلى تحليل المشاعر أو العواطف على أنّه: عملية تصنيف تهدف إلى تحديد ما إذا كانت وثيقة ما أو نص مكتوب تُعبّر عن رأي إيجابي أو سلبي حول كائن معين (على سبيل المثال، موضوع أو مُنتَج أو شَخص). وتتعامل هذه العملية مع كلّ وثيقة وحدةَ

معلومات أساسية. ويُشار إلى هذه العملية على أنّها «تصنيف لمستوى العاطفة في الوثيقة». كما تمت دراسة المشاكل الأكثر دقةً لتحديد مستوى العاطفة لكلّ جُملة. وهناك أسلوبان أساسيّان مُستخدمان في تحليل المشاعر أو العواطف وهما: أولاً، المصنّفات القائمة على القواعد، حيث يتم تطبيق القواعد المستمدة من دراسة القواعد اللغوية على عملية تحليل العواطف، وثانياً، المصنّفات المعتمدة على مبدأ التعلم الآلي، حيث تُستخدم خوارزميات التعلّم الآليّة الإحصائية للتعرّف على إشارات المشاعر تلقائياً.

إنّ مُعظم أنظمة تحليل العواطف المستخدمة حالياً مُصمّمة للتعامل مع اللغة الإنجليزية، لكن يُذكر بأن هناك بعض الجُهود التي بُذلت للعمل على لغاتٍ أخرى. وتمتاز اللغة العربيّة بأنّها غنيّة جداً بوظائف الصّرف والنّحو، وهي تُعدّ واحدة من أغنى اللغات من حيث قواعد الصّرف. من جهة أخرى، بيّنت الدراسة أهمية توافُر المتون المشروحة لغرض التدريب أو التمرن عليها واختبارها لتطوير أنظمة التعرّف على المشاعر أو العواطف.

المباحث الرئيسية للدراسة

أولاً: خصائص اللغة العربيّة

تضمن البحث معلومات عامة عن خصائص اللغة العربية

ثانياً: المتون والمعاجم

إن توافر المُتون المشروحة لأغراض التدريب والاختبار مُهم جداً لتحقيق تقدم في أنظمة التعرف على العواطف والأحاسيس. لكنّ جمع مثل هذه البيانات (وخاصة الشُروح) قد يتطلّب توفير أعداد كبيرة من العاملين (عمالة مكثّفة). ولحُسن الحظ، قام عدد من الباحثين بتطوير مُتون خاصة بتحليل العواطف العربية وإصدارها، ولقد استعرضت هذه الورقة بعضاً منها. وتراوحت تلك المتون بين المصنِّفات القائمة على القواعد، حيث اعتمدت على تطبيق القواعد المستمدة من قواعد اللغة العربيّة على عملية تحليل العواطف، والمصنِّفات المعتمِدة على مبدأ التعلّم الآلي، حيث استخدمت فيها خوارزميات التعلّم الآليّة الإحصائية للتعرّف على إشارات المشاعر أو دلالاتها بشكل تلقائي.

ثالثاً: أنظمة تحليل الذاتية والعواطف وأساليبها في اللغة العربيّة

يتناول هذا الجزء من الدراسة مراجعة الأساليب المختلفة لتحليل الذاتية في اللغة العربيّة، وتحليل العواطف التي اقتُرحت في الأدبيات ذات الصّلة، حيث ناقشت الدراسة أولاً أساليب الاختيار / الاستخراج المستقلة عن اللغة العربيّة التي تمّ تطبيقها على اللغة العربيّة. ثمّ ناقشت الأنظمة التي تستخدم أساليب (IR) القياسية (مثل: TF*IDF) والمصنِّفات الهجينة). وأخيراً، ناقشت الدراسة – ضمن هذا المبحث – الأنظمة التي قامت بتوظيف المزايا الخاصة باللغة العربيّة وتلك المصمّمة خصيصاً لاستخراج أصحاب الرأي أو التعرف عليهم. وفيما يلي تفصيل لكل أسلوب من هذه الأساليب.

منهجية الاستخراج والاختيار المستقل عن اللغة

استعرض هذا المبحث العديد من الطُرق والأساليب لاختيار مثل تلك الميزات والخصائص واستخراجها، من ضمنها: (١) الخوارزميات الجينية المرجّحة، (٢) تصنيف الخصائص أو المزايا، (٣) الأساليب المعتمدة على القواعد النّحوية المحليّة، (٤) الخصائص أو المزايا الموضعيّة، وأخيراً (٥) أساليب (الأنوية) المشتركة للكلمة. ومن بين هذه الطُرق والأساليب استوقفت الدراسة اثنتين منها فقط، وهما: الخوارزميات الجينية المرجّحة والأساليب المعتمدة على القواعد النّحوية المحليّة.

قيّمت الورقة النّظام الأول المقترح (الخوارزميات الجينية (الوراثية) المرجّحة) استناداً الى اختبار قياسي مكون من ١٠٠٠ رأي إيجابي و ١٠٠٠ رأي سلبي تجاه أحد الأفلام المعروضة على الشاشة. في حين يستند النظام الثاني المقترح (المعتمد على القواعد النحوية المحليّة) إلى تحديد المفردات الرئيسية للنطاق من خلال البحث عن الكلمات المتكررة بكثرة في متن قائمة الأخبار المالية، لكنها تكون نادرة نسبياً في القائمة العامة. وباستخدام السياق المحيط بهذه الكلمات، يقوم النظام ببناء قواعد محلية لاستخراج عبارات تحمل المشاعر أو العواطف. ويُستخدم هذا الأسلوب الأخير على اللغات العربيّة والإنجليزية والصّينية. وقُيّم النظام يدوياً حيث حقّق مُعدلات دقة تتراوح بين ٦٠-٧٥٪ لاستخراج العبارات التي تدلّ على المشاعر أو العواطف. لكن الأهم أنّه يمكن استخدام النظام المقترح لاستخراج عبارات المشاعر أو العواطف في المجال المالي لأيّ لغة.

أسلوب المُصنِّفات القياسية (IR) والهجينة

ناقشت الدراسة ضمن هذا الجزء الأنظمة المعتمدة على استخدام أساليب (IR) القياسية (مثل: TF*IDF) والمصنِّفات الهجينة)، حيث استعرضت الدراسة العديد من الأعمال التي قام بها الباحثون في استخدام تطبيقات هذا النظام، وكان من أبرز نتائجها:

تقديم نظام لتحليل المشاعر المطبّق على بعض المراجع التجارية العربيّة، مع التركيز على

الهدف المتمثّل في بناء مُحرّك بحث على الشّابكة يمكنه أن يُعلّق تلقائياً على الصفحات المسترجعة مع إعطاء علامة أو مجموع النقاط لكل منها من حيث محتواه من العواطف، ويحتوي هذا النظام على عدة مكونات، حيث يُصنّف المكوّن الأول ما إذا كانت الصفحة المنشورة على الشابكة مرجعية أم لا. ثمّ يقوم النظام بتحليل الوثيقة من حيث محتواها من العواطف، ويقوم بإنشاء مُعجم عربي على أساس نفس أداة التسجيل لاستخدامه في الكشف عن محتوى العواطف. أما المكوّن الأخير للنظام، فهو مُصمّم لتزويد مُحرّك البحث بتقدير علامة للعواطف الواردة في الوثيقة خلال عملية البحث.

اقترُح نظام لتصنيف محتوى العواطف على مستوى الوثيقة، وطُبّقت مُصنّفات مُختلفة قائمة على أساس هذا المبدأ، بطريقة تتابعية، حيث استُخدم في البداية المصنف المبني على المعجم لتقدير محتوى الوثيقة المعنيّة من العواطف، بناءً على مجموع كلّ الكلمات والعبارات التي تتضمن تعبيراً عن الرأي الواردة في تلك الوثيقة. وقد تضمن هذا النظام متناً مكوّناً من ١١٣٤ وثيقة جُمعت من مجالات مختلفة (مثل التعليم والسياسة والرياضة)، حيث أظهرت نتائج التحليل عليها وجود ٦٣٥ وثيقة إيجابية (تضمنت ٤٣٧٥ جُملة إيجابية) و ٥٠٨ وثيقة سلبية (تضمنت ٤١١٨ جُملة سلبية).

أسلوب سمات خصوصية اللغة العربيّة ووسائل التواصل الاجتماعي والخصوصية الأدبية

وهناك تقنيات أخرى تستخدم السمات اللغوية للغة العربيّة في إجراء تحليل العواطف، من خلال تحليل التركيب النحوي والخصائص الصّرفية الخاصة للغة العربيّة. واتبع بعضُهم تصنيفاً للعواطف على مُستوى الجُملة العربيّة، آخذين بالاعتبار نهجين مُختلفين: النهج النّحوي والنهج الدّلالي. ويعتمد النهج الأول على قواعد النّحو العربي ويجمع تراكيب الجُملة الاسمية والجُملة الفعلية في صيغة عامة واحدة بناءً على الفكرة/ الفعل. وفي هذا الأسلوب يُمثّل (الفاعل) في كل من الجُمل الاسمية والفعلية (الشّخص الذي يقوم بأداء الدّور)، في حين تُمثّل الأفعال (الأداء أو الدّور)، فتمّت تسمية علامات للفعل والفاعل في الجُمل ومن ثمّ تُستخدم هذه العلامات

ميزاتٍ أو سِماتٍ. ويتكوّن مُتجه العناصر الخاصة بهذا الأسلوب من الأبعاد التالية:

نوع الجُملة (اسمية أو فعلية)، فاعل (الشّخص)، فعل (دور)، كائن، صفة، نوع الضمير والكلمات أو العبارات الانتقالية (نوع الكلمة أو العبارة التي تربط الجُملة الحالية بالجملة السابقة)، وجاذبية الكلمة: (إيجابية، سلبية، مُحايدة) وفئة الجُملة. وتضمّن هذا المبحث مُقارنة بين الأنظمة السابقة لتحليل العواطف، حيث بيّنت الدراسة أنّ لكل نظام مزايا وعيوبا من حيث سهولة الاستخدام والكُلفة وارتباطه بخصائص اللغة العربيّة أو استقلاله عنها وكذلك أنواع المتون المستخدمة فيها.

أسلوب استخراج (التعرُف على) صاحب (حامل) الرأي

اقتُرحت عدة أساليب مُختلفة لاستخراج صاحب الرأي في اللغة العربيّة، وقد استندت تلك الأساليب إلى كلّ من مُطابقة النمط والتعلم الآلي. وتوصّلت هذه الأساليب إلى استخراج ثلاثة أنواع مُختلفة من أصحاب الرأي؛ النوع الأول: هو صاحب الرأي تجاه أحداث الكلام، ويُعرّف بأنه بيان شخصي قيل مُباشرةً من شخص ما أو ادّعى شخص ما قولَه. وبهذه الطريقة، جُمع – ضمن هذا النوع – بين حدث الكلام المباشر وحدث الكلام غير المباشر. ويُعرّف النوع الثاني: بأنه مُرتبط بحامل رأي يُعبّر عن عواطفه تجاه موضوع رأي مُعين. ويعرّف النوع الثالث: بأنه يتعلق بالعناصر الذاتية التّعبيرية (مثل: العاطفة والسُخرية) المعبّر عنها بشكل ضمني. ومن المؤكّد أن النوع الثالث هو الأصعب بسبب اعتماده على معنى الكلمات بدلاً من الهياكل أو التراكيب اللُغوية، في حين يعتمد النوع الأول لاستخراج أصحاب الرأي على مُطابقة النّمط، حيث اتُّبعت القاعدة التالية لاستخراج صاحب الرأي: يُسترجع صاحب الرأي إذا كان النصّ يحتوي على بيان شخصي أو كيان مُحدد ويحتوي على بيان مُصنّف على أنه موضوعي أو غير موضوعي باستخدام مُصنّف عالي الدقة.

يعتمد الأسلوب الأول على مطابقة النمط، ويستند الأسلوبان الثاني والثالث إلى التعلّم الآلي. وقد تعامل الباحثون مع مُشكلة حامل الرأي مُشكلةَ تصنيف، حيث تُصنّف كل كلمة في

المتن على أنها «بداية حامل الرأي (ب- حامل)»، أو «داخل حامل الرأي (د- حامل)» أو «غير حامل لرأي». واستُخدم نموذج تمييز الحقول العشوائية المشروطة (CRF) في عملية التّصنيف. وقد قام الباحثون ببناء المصنّف المذكور اعتماداً على مجموعة من الخصائص أو السّمات المعجمية والصّرفية، والدّلالية.

وأظهرت النتائج التجريبية على أسلوب «استخراج صاحب الرأي العربي» أنّ أسلوب التعلّم الآلي المستند إلى نموذج تمييز الحقول العشوائية المشروطة (CRF) حقّق نتائج أفضل من أسلوب مُطابقة الأنماط. وحسب الباحثين فإنه حقّق نسبة ٥٢ , ٨٥٪ من الدّقة، و ٤٩ , ٣٩٪ من الاستدعاء، وحصل على نسبة ٠٣ , ٥٤٪ لمقياس F. وقد عزا الباحثون ضعف أداء النظام إلى ضعف أداء أدوات البرمجة اللغوية العصبية العربيّة مُقارنة بتلك المستخدمة في اللغة الإنجليزية فضلاً عن عدم وجود مُحلّل مُعجمي.

رابعاً: الخاتمة

ناقشت هذه الدراسة المسحيّة الأساليب المختلفة لبناء أنظمة تحليل الذات والعواطف في اللغة العربيّة، واستعرضت الموارد المتاحة لتحليل العواطف في اللغة العربيّة. واقترحت الدراسة اتباع طريقة دقيقة في بناء أنظمة تحليل العواطف العربيّة لأنّ ذلك من شأنه أن يساهم في الاستفادة، ليس فقط من سمة الاستقلالية عن اللغة، ولكن أيضاً من سمات خُصوصيّة اللغة العربيّة؛ واستغلال أوسع مدى للمصادر والمعاجم المتخصصة. كما أنها ستُساهم في رفع سويّة سمات خُصوصيّة شبكات التواصل الاجتماعي ووسائل الإعلام والسمات ذات الخُصوصية الأدبيّة. وعلى الرغّم من أن كُلفة بناء موارد مُصمّمة خِصيصاً للغة العربيّة والحصول على سِمات مُعينة، سوف يؤدي سلوك هذا الطريق إلى إنتاج عالي الأداء وسوف يُضيف نظرة ثاقبة ومُثيرة للاهتمام لمهمة التّصنيف. أما البدائل الممكنة عن هذه الطريقة فتتمثل في نقل المعرفة المتعلّقة بالعواطف من اللغة الإنجليزية إلى اللغة العربيّة أو استخدام أساليب مُستقلة لُغوياً.

(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 7, No. 5, 2016

Year of publication: 2016

Semantic Feature Based Arabic Opinion Mining Using Ontology

Abdullah M. Alkadri, Abeer M. ElKorany

التّنقيب عن الرأي العربي القائم على السِمات الدلالية باستخدام الأنطولوجيا

المقدمة

مع التّزايد الكبير في عَدد موارد الرأي المنشورة على الشّابكة، بات يُعد التحليل والاستخراج التلقائي للمَعرفة من تلك الموارد أو المراجع أمراً مُهماً جداً، لكن يصعُب أداء مثل هذه المهمة يدوياً. إنّ التنقيب أو البحث عن الرأي هو أحد أشكال المعرفة المتعلّقة بالتنقيب داخل النصوص، ويقوم بأداء مثل هذه المهمة تلقائياً. وقد ركّزت معظم البحوث التي أجريت في هذا المجال على النُصوص الإنجليزية مع وجود عدد مَحدود جداً من الأبحاث التي أُجريت على اللغة العربيّة، وتُعزى هذه النُدرة لوجود الكثير من العقبات في اللغة العربيّة. وتهدف هذه الورقة إلى تطوير إطار إبداعي لاستعراض الرأي القائم على السِمات الدلالية للموارد العَربيّة. ويستفيد هذا الإطار من الدلالات اللغوية للأنطولوجيا والمعاجم في تحديد سِمات الرأي ودرجة قُطبيّتها. لقد أظهرت التجارب بأن الإطار المقترح في هذه الدراسة حقق مُستوىً جيّدا من الأداء مقارنة ببيانات الاختبار التي جُمعت يدوياً.

وتناولت الدراسة المباحث الرئيسية الآتية:

أولاً: بناء المُعجم والأنطولوجيا

بناء الأنطولوجيا

في هذا الجُزء، تستعرض الدراسة عملية بناء الأنطولوجيا. ولغرض بناء أنطولوجيا مُتخصّصة بمجال مُعيّن يمكن تطبيق أحد أسلوبين: (١) استخدام أنطولوجيا قائمة، (٢) بناء أنطولوجيا جديدة من الصّفر. لقد بدأ عملنا بالبحث عن أنطولوجيا معروفة مُتخصّصة في هذا المجال باللغة العربيّة، لكن تعذّر علينا العُثور عليها، وبالمقابل، وجدنا أنطولوجيا إنجليزية لتقييم الفنادق (Hontology) التي قمنا بترجمتها يدوياً للغة العَربيّة. ونظراً لنقص المعرفة والمعلومات بالمجال المقصود وبسبب الغموض في التسلسل الهرمي للأنطولوجيا فقد واجهناها من خلال مرحلة التنقيح.

بناء مُعجم واسع النطاق للرأي العَربي

تعتمد معظم أساليب استخراج الرأي على مُعجم الرأي، مثل المعجم الإنجليزي (SentiWordnet English) ومعجم (MPQA) لتحديد قُطبية الكلمة، ومن أجل الحصول على دقة أعلى، يوصى باستخدام مُعجم رأي باللغة العربيّة كبير الحجم. ولقد ظهرت مؤخراً بعض معاجم الرأي العربيّة، لكن توافر معجم الرأي العربي على نطاق واسع لا يزال محدوداً وغير متاح. وبالنظر إلى محدودية معاجم المشاعر العربيّة، فإننا نقترح معالجة هذه المشكلة من خلال تطوير معجم للرأي العربي وبنائه على نطاق واسع (ArOpL).

ثانياً: الإطار المُقترح للتنقيب عن الرأي العربي القائم على السِمات الدلالية

في هذا الجزء من الدراسة، تمّ تقديم شَرح للمنهجية المقترحة، وبنية هذا النظام. ويتكون هذا الإطار من خمسة عناصر رئيسية، هي: المعالجة المسبقة، وتحديد السمات الدلالية، وتحديد القُطبية، وتحديد قطبية السمات واستخراج الرأي. وفيما يلي وصف لهذه العناصر بالتفصيل.

المُعالجة المسبقة: يلزم تطبيق العديد من تقنيات معالجة اللغة الطبيعية (NLP) على قائمة البيانات الموجودة لضمان تنقية البيانات وإزالة الشذوذ الذي قد يؤثر على دقة النظام.

تحديد السِمات الدلالية: استُخدمت أنطولوجيا المجال (النطاق) لتحديد السمات الدلالية

في تقييمات الفندق. وبعد المعالجة المسبقة لتقييمات الفندق وتحديد الأسماء في التقييمات باستخدام أدوات وسم أقسام الكلام (POS taggers) تمّ اتباع ثلاث خطوات لاحقة للعُثور على السمة المقابلة باستخدام أنطولوجيا المجال أو النطاق.

تحديد القُطبيّة: من أجل تحديد القُطبيّة، يُعد وجود قائمة من كلمات الرأي ضرورياً، فكلمات الرأي هي كلمات تعبّر عن المشاعر الإيجابية أو السلبية. وهكذا، استخدمنا المعجم الكبير الجديد (ArOpL). وبعبارة أخرى، لكلّ تقييم حددت طريقتنا درجات +١ و -١ للكلمات الإيجابية والسّلبية على التوالي.

تحديد قُطبية السمات (تشكيل قائمة المكونات أو الحُقول المُترابطة): باستخدام السمات المستخرجة وقوائم الكلمات الإيجابية والسلبية الناتجة عن المراحل السابقة، قمنا بتحديد اتجاه الرأي المعبّر عنه لكل سمة. ومن هذه الخطوة أمكن الحصول على قائمة من الصفوف التي تحتوي على السمات وأقطابها.

التنقيب أو استخراج الرأي: يتم الحصول على القطبية العالمية للتقييم من خلال تحديد غالبية السمات المستقطبة التي حددها نظامنا بالفعل. فإذا كانت السمات الرئيسية مستقطبة على أساس أنها إيجابية، فإن القطبية العالمية لها تعدّ إيجابية. وبالمثل، في حال إذا كانت السمات الرئيسية مستقطبة على أنها سلبية، وبخلاف ذلك فإن القُطبية العالمية محايدة.

ثالثاً: التّجربة والمناقشة

الإعداد للتجربة: نظرا لعدم وجود بيانات موسومة في مجال الفنادق العربيّة، فقد جمعنا تقييمات الاختبار يدوياً من مجموعة متنوعة من المواقع ذات الصلة بمجال الفنادق، حيث تحسسنا هذه التقييمات من بلدان مختلفة ومن ثلاثة مواقع هي: (www.tripadvisor.com) (http://www.booking.com) و (http://www.agoda.com) للعثور على البيانات ذات الصلة بمجال الفنادق. وبلغ إجمالي عدد التقييمات أو الآراء التي استُخدمت ٨٩٠ رأياً، حيث استُخدم ٦٩٠ تقييماً من أجل تمديد الأنطولوجيا أما التقييمات الباقية التي استخدمت للتجارب

والبالغة ٢٠٠ تقييم، فقد كان نصفها سلبيا، والباقي إيجابيا. ثم قمنا بوضع علامات (وسم) التقييمات يدوياً لجمع نتائج خط الأساس المستخدمة لتقييم المنهجية المقترحة.

المناقشة: قُسّمت النتائج إلى فئتين مختلفتين: متوسط دقة كلّ من تحديد قطبية السمات، وتصنيف استخراج الرأي لكامل الوثيقة بالنسبة للتقييمات الإيجابية والسّلبية على حد سواء. وأظهرت النتائج أن أفضل عمليات تحديد لقطبية السمات المتعلقة باستخراج الرأي باستخدام طريقة N-GRAM كانت عند مستوى (N-GRAM=٤) مع مستوى دقة بلغ ٧٢٪ ، في التعليقات الإيجابية و ٦٣٪ في التعليقات السّلبية. وهذا يعني أن القطبية القائمة على السّمة تحسب باستخدام ٤ كلمات قبل السمة و ٤ كلمات بعد السمة في تقييم المستخدمين قد حققت درجة دقة جيدة. وفي الواقع، كانت أسوأ النتائج التي تم الحصول عليها عند مستوى (N-GRAM=١) وبدرجة دقة بلغت ٥٤٪ في التعليقات الإيجابية و ٤٥٪ في التعليقات السلبية. في حين تم الحصول على أفضل نتائج التنقيب أو استخراج الرأي عند مستوى (N-GRAM=٤) مع درجة دقة بلغت ١٠٠٪ في التعليقات الإيجابية و ٩١٪ في التعليقات السلبية.

وأخيراً، من بين الطُرق الثلاث المقترحة حققت طريقة (N-GRAM) أفضل النتائج لكل من عملية تحديد قُطبية السمات، والتنقيب عن الرأي بين المستخدمين المتحدثين بالعربيّة، حيث حصلت على مستوى دقة ٥ , ٦٧٪ و ٥ , ٩٥٪ لجميع التقييمات، على التوالي.

التوصيات

يُعدّ التّنقيب عن الرأي العربي مُشكلة صَعبة؛ فهي تُعنى بتحليل الآراء التي تظهر في تقييمات المستخدِمين، وتحديد إذا كانت هذه الآراء إيجابية أو سلبية. في هذه الورقة، اقتُرحت منهجية للتنقيب عن الرأي العربي القائم على السمات، بحيث تمر هذه المنهجية بخمس مراحل مُختلفة: الأنطولوجيا وصناعة المعجم ؛ وتحديد السمات الدلالية؛ وتحديد القُطبية، وتحديد قُطبية السمات؛ وأخيراً التنقيب عن الرأي.

وعلى الرغم من جميع المزايا والإمكانات التي تتمتع بها المنهجية المقترحة، إلا أنها تواجه

العديد من القيود التي يمكن تحسينها في المستقبل. أولا: يمكن تحسين المنهجية المقترحة من خلال إدخال تقنيات استخراج الرأي القائمة على التّعلّم الآلي؛ ثانيا: بما أن علم الأنطولوجيا الحالي ساكن، والمعرفة الممثلة فيه لا تكفي، سيكون من المثير للاهتمام بناء أنطولوجيا شبه تلقائية تعتمد على تقنيات التعلم بالأنتولوجيا من خلال تقييمات المستخدمين. وأخيراً، توصي الدراسة بوضع خُطط مُستقبلية لتطبيق المنهجية المقترحة في نطاق آخر مثل تقييم المنتجات.

2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)

Year of publication: 2013

Arabic Sentiment Analysis: Lexicon-based and Corpus-based

,Nawaf A. Abdulla, Nizar A. Ahmed, Mohammed A. Shehab Mahmoud Al-Ayyoub

تحليل المَشاعِر العربيّة القائِمة على المُفردات والمَتن

المقدمة

لقد أدّى نُشوء تقنية (Web 2.0) إلى إنتاج كمّ هائل من البيانات الخام من خلال تمكين مُستخدمي الشّابكة من نشر آرائهم وتقييماتهم وتعليقاتهم على الشّبكة. إلاّ أنّ مُعالجة هذه البيانات الخام لاستخراج المعلومات المفيدة منها يمكن أن تكون مُهمة صَعبة جداً. ومن بين الأمثلة على المعلومات المهمة التي يمكن استخلاصها تلقائياً من مُشاركات المستخدمين وتعليقاتهم تحديد آراء المستخدمين حول مختلف القضايا والأحداث والخدمات والمنتجات، إلخ. وقد تمت دراسة مُشكلة تحليل المشاعر أو العواطف بشكل جيد على اللغة الإنجليزية، حيث تمّ استنباط منهجيتين رئيسيتين: الأولى تستند إلى المتن اللغوي، والأخرى تستند إلى المفردات المعجميّة، وقد استعانت هذه الورقة بكلا المنهجيّتين في معالجة تحليل المشاعر أو العواطف في اللغة العربيّة.

المُشكلة البحثية

نظراً لمحدودية البيانات والمعاجم العربيّة المتاحة الخاصة بتحليل المشاعر أو العواطف، بدأت هذه الوَرقة من بناء قائمة بيانات مشروحَة يدوياً تنقل القارئ لاحقاً خلال الخُطوات

التفصيلية لبناء المعجم. وأُجريَت التّجارب خلال المراحل المختلفة من هذه العملية لمراقبة التّحسينات التي تحقّقت على دقة النظام ومقارنتها مع المنهجيّة القائمة على المتْن.

وفيما يلي المباحث الرئيسية التي ناقشتها الورقة:

أولاً: المنهجية أو الأسلوب

في هذا الجُزء من الدراسة ناقشنا المنهجيّتين كلتيهما؛ المنهجيّة القائمة على المتن اللغوي، والمنهجيّة القائمة على المفردات المعجميّة. اعتمدت الخطوة الأولى على جمع قوائم المرادفات وشَرحها، وهي خطوة مكلفة من حيث الوقت والجهد، حيث استخدمت قائمة المرادفات لبناء نموذج تصنيف لأداة قائمة على المتن، بالإضافة الى اختبار الأداة القائمة على المعجم. وفيما يلي قائمة التحديات التي واجهت بناء المعجم.

قائمة البيانات والمعالجة المسبقة: يمكن أن يكون لنوع الوثائق التي تتضمن مجموعة أو قائمة من البيانات (سواءً كانت مُقارنات أو مُراجعات أو تَعليقات أو تَغريدات أو غير ذلك) تأثير مُهم على تحليل المشاعر أو العواطف، حيث يحكم حجم المستند واللغة المستخدمة فيه (إلى أي حدٍ هي رسمية أو اصطلاحية، والموارد الثقافية المستخدمة، وفيما إذا كانت تستخدم السخرية، وهل هي لغة واحدة أو خليط من عدة لغات، الخ). وقررنا العمل على التّغريدات نظراً لشَعبيّة مواقع شبكات التّواصل الاجتماعي، والقُدرة على استخراج الكثير من المعلومات المفيدة منها (من آراء ومشاركات الناس المنشورة على مثل هذه المواقع). وباستخدام "داب التّغريدات»، تمكّنا من جمع ٢٠٠٠ تغريدة مُعلّمة (مُصنّفة) (١٠٠٠ تغريدة إيجابية و ١٠٠٠ تغريدة سلبية) تتعلق بمواضيع مختلفة مثل: السياسة والفنون. وتشتمل هذه التغريدات على الآراء المكتوبة بالعربيّة القياسيّة الحديثة واللّهجة الأردنية المحلّية.

بناء المعجم: نظرا لطبيعتها المعقّدة خصصنا معظم جهودنا لبناء المعجم يدويا وتعزيزه، حيث بدأنا بـ ٣٠٠ كلمة جذرية مأخوذة من موقع (SentiStrength). وتُرجمت هذه الكلمات أولا إلى اللغة العربيّة باستخدام قاموس إنجليزي-عربي، ثُمّ أعطيت الكلمات الموجبة رقم

قطبية (+١) في حين أُعطيت الكلمات السلبية رقم قُطبية (-١). ثمّ قُمنا بتجميعها معاً في مُعجم واحد اختصاراً للوقت على أداة بحثنا. ولتَحسين أداء الأداة القائمة على المعجم، تمّ تطبيق عدة مُلحقات (توسعات) على المعجم. وأخيراً، تم توسيع المعجم الناتج من ٣٠٠ كلِمة إلى ٣٤٧٩ كلمة، تتكون من ١٢٦٢ كلمة إيجابية و ٢٢١٧ كلمة سلبية.

تصميم الأداة والتنفيذ: تم تصميم العديد من الخوارزميات والأدوات للتعامل مع تحليل المشاعر غير الخاضعة للرقابة. ويستخدم العديد من الباحثين تقنية قائمة على مُفردات المعجم لتحليل المشاعر أو العواطف، لتجنب الطريقة اليدوية المكلفة، للتعليق على مجموعة بيانات. وبناء على ذلك، بنينا في منهجيتنا المتبعة هنا أداةً للعثور على التوجهات العاطفية للنّص العربي، بحيث يكون لكل كلمة أو مُفردة في المعجم وزن يمثل قطبية موجبة أو سالبة (+١ و -١ للكلمات الإيجابية والسلبية، على التوالي). وأخيراً، وبعد استخراج قطبية كلّ كلمة من المعجم العربي، تقوم الأداة بتجميع إجمالي الأوزان وتحدد قطبية النص المدخل بأكمله.

ثانياً: النتائج

الهدف من التجارب هو المقارنة بين دقة كل من الأسلوبين في تحليل المشاعر (الأسلوب القائم على المتن والأسلوب القائم على المعجم)، وإظهار كيف تتحسّن دقة الأداة القائمة على المعجم مع إضافة المزيد من الكلمات إلى المعجم. ولتقييم دِقة كل أداة، اعتمدنا على ثلاثة من أكثر مقاييس الدِقة استخداماً في الأدبيات ذات الصّلة وهي الدقة والاسترجاع والإحكام (الضبط)، صُممت معادلات رياضية لكلٍ منها. كما أُجريت ثلاث تجارب مختلفة باستخدام أربعة مُصنّفات شهيرة وهي: (SVM، NB ، D-Tree ، KNN). وأظهرت نتائج هذه التجارب الثلاث أنه كلما كان حجم المعجم أكبر كلما كانت النتائج أفضل. من جهة أخرى، لوحظ من النتائج بأن فجوات الدقة تتقلص أكثر مع ازدياد حجم المعاجم، لكن يجب الانتباه في الختام إلى أنه يمكننا التكهن بأن حجم العمل (أي حجم المعجم) المطلوب لكل زيادة في درجة دقة الأداة القائمة على المعجم ينمو بسرعة، مما يعني أنه قد لا يكون الخيار الأفضل دائماً هو توسيع المعجم

من أجل تحسين الدقة. فقد تكون المعالجة الفضلى والطرق الأكثر ذكاء لحساب قطبية النص (استنادا إلى استقطاب الأفراد) ذات تأثير أكبر على درجة الدقة.

ثالثاً: الاستنتاج والتّوصيات

ناقشت هذه الورقة منهجيتين لتحليل المشاعر: الأولى تستند إلى المتن اللغوي، والأخرى تستند إلى المفردات المعجميّة. وبالنظر إلى المحدوديّة والنُدرة في قائمة البيانات العربية والمعاجم المتاحة للجمهور لأجل تحليل المشاعر أو العواطف، فقد ناقشت هذه الورقة بناء قوائم بيانات مشروحة يدوياً تنقل القارئ إلى الخطوات التفصيلية لبناء المعجم. ولوحظ من نتائج التجارب التي أجرتها الدراسة أنّ أداة التحليل القائمة على المتن الذي يستخدم خوارزمية **SVM** لتصنيف قوائم البيانات المستمدة من الجذور الضعيفة (الخَفيفة) تُعطي أعلى درجة مُمكنة من الدقة. وعلاوةً على ذلك، لوحظ أنه مع زيادة حجم المعجم، تتحسّن دِقّة أداة التحليل القائمة على المعجم. وبينما هذه النتائج مَعروفة أو مُتوقّعة، فإن الملاحظة الأخرى جديرة بالتفكير، وهي أنّ مقدار العمل (أي حجم المعجم) المطلوب لكل زيادة في دقة الأداة القائمة على المعجم يزداد بسرعة أيضاً.

وتمثل هذه الدراسة الأساس لعمل مستقبلي يُخطّط لتوسيع قائمة البيانات جنباً إلى جنب مع إضافة حالة قطبية ثالثة (الطبقة المحايدة). أضف إلى ذلك، زيادة حجم المعجم ستُوفّر نتائج أفضل بكثير، خُصوصاً عند إضافة قوة إلى قُطبية الكلمات بحيث تتراوح درجة القُطبية ما بين -٥ إلى +٥، وقد تأتي بنتائج أكثر دقة.

Advances in Soft Computing and Its Applications. MICAI 2013. Lecture Notes in Computer Science, vol 8266. Springer, Berlin, Heidelberg

Year of publication: 2013

Finding Opinion Strength Using Rule-Based Parsing for Arabic Sentiment Analysis

Shereen Oraby, Yasser El-Sonbaty, Mohamad Abou El-Nasr

قياس قوة الآراء باستخدام التحليل اللغوي المعتمد على القواعد للمواقف باللغة العربيّة

ملخص البحث

مع زيادة الاهتمام ببحوث تحليل المواقف وتنامي محتوى الإنترنت الحافل بالآراء، من المهم أن تركِّز الأبحاث على تحليل النصوص من مختلف المجالات واللغات. يتحرى هذا البحث مشاكل تحليل المواقف وقياس قوة الآراء باستخدام منهج معتمد على القواعد يناسب اللغة العربيّة. وتراعي مقاربتنا الخصائص الفريدة للغة العربيّة والضرورية من أجل تقسيم النص بحسب القواعد التركيبية والسماح بإجراء تحليل أدق للوحدات اللغوية الحاملة للآراء. وباستخدام مجموعة مخصصة من مفردات المواقف ومؤشرات الآراء، نطرح منهجية قائمة على القواعد لاستخراج عبارات الرأي، تعقبها طريقة لتصنيف الآراء المحللة ومقياس لقوة الآراء في النصوص التي نتناولها بالتحليل. إن الطريقة المقترحة في بحثنا تُظهر الفرص التي يقدمها نظام بسيط لتصنيف الآراء يتصف بقابلية التوسعة، بما يناسب اللغات الثرية بالقواعد الصرفية مثل اللغة العربية.

مقدمة

أصبح تحليل الآراء والمواقف من مواضيع البحث الرئيسية في السنوات الأخيرة، ولا سيما بسبب التطبيقات الضخمة والفرص الصناعية الكامنة في جمع البيانات الإحصائية حول

استقطاب الآراء في سياقات متنوعة، منها الاجتماعي والسياسي والتجاري. وثمة حاجة متزايدة إلى أنظمة تصنيف تلقائية تتسم بالكفاءة والموثوقية وقابلية التوسع، وذلك من أجل استغلال الكمّ الهائل من بيانات الآراء المتوفرة على الإنترنت في التصنيف والتحليل.

نلاحظ محدودية البحوث المجراة على اللغات الثرية بالقواعد الصرفية مثل اللغة العربيّة، ولا سيما فيما يتعلق بالمقاربات المعتمدة على القواعد، التي تحاول تحديد أنماط تركيبية باللغة للمساعدة في مهام التحليل. فتعقيد البنية التركيبية والنحوية في اللغات الثرية بالقواعد الصرفية يجعل مهمة نمذجة اللغة وتمثيلها تحديًا لأنظمة معالجة اللغات الطبيعية. ففي اللغة العربيّة، توجد حرية كبيرة في ترتيب الكلمات والبناء التركيبي، حيث إن القواعد الصرفية نفسها تحدد كثيرًا من جوانب تركيب اللغة.

تناولت دراسات سابقة كثيرة تصنيف المواقف تصنيفًا ثنائيًا تقليديًا إلى «إيجابية» و«سلبية». أما بحثنا فيركّز على تحليل المواقف باللغة العربيّة من خلال مقاربة معتمدة على القواعد باستخدام مفردات المواقف وقوائم الكلمات المقيِّدة للمعنى من أجل تحديد موقف النصوص الموجودة على مكنز لمراجعات الأفلام على الإنترنت. وقد اخترنا هذه المجموعة من البيانات نظرًا لمحتواها الذي يصطبغ بالآراء الشخصية ومقياس تقييم المراجعات. يحتوي المكنز على ٤٨٣ مراجعة عربيّة للأفلام، منها ٢٥٠ مراجعة إيجابية و٢٣٣ مراجعة سلبية، مُجمعة من ١٥ موقعًا عربيًا مختلفًا على شبكة الإنترنت. وتتكون كل مراجعة من نص يعرض تفاصيل الفيلم وتعليق صاحب المراجعة. وكانت أغلب المراجعات مصحوبة بتقييم رقمي للفيلم دعمًا للمراجعة المكتوبة. وقمنا بمقارنة النصوص (تحديدًا المراجعات) المختارة بناءً على قوة الموقف (في مقابل التقييمات التي أعطاها أصحاب المراجعات) وبالتالي حصلنا على مقياس أكثر تفصيلًا لقوة المواقف.

بالنسبة إلى تصميم النظام، قمنا في البداية بتحليل نصوص المكنز وقوائم الكلمات المقيّدة للمعنى باستخدام مجموعة من قواعد النحو، وقسمنا النصوص إلى أنماط مكونة من العبارات والكلمات المقيِّدة للمعنى التي ستُستخدم في مقارنتها مع مجموعات الكلمات الاستقطابية.

وتتكون مجموعات الكلمات الاستقطابية من أربع مجموعات من مفردات المواقف (مجموعات الكلمات القوية والضعيفة لكل من التصنيف الإيجابي والتصنيف السلبي)، التي ستُستخدم في المقارنة مع الكلمات في النصوص المحلَّلة بالفعل بناءً على القواعد في مرحلة حساب نقاط الآراء. يُعطى كل نص نقاطا إيجابية ونقاطا سلبية بحسب عدد وحدات الاستقطاب والكلمات المقيِّدة للمعنى في كل جملة بالنص، ثم يُعطى كل نص تصنيفًا، بناءً على مقياس لنسبة الوحدات الاستقطابية به، ما يُسفر عن تصنيف القوة لكل نص من النصوص. بالتالي، يمكن مقارنة تصنيف القوة بتقييم المراجعة الأصلي، وهو ما يمكن استخدامه في تعيين دقة النظام وغيرها من مقاييس التقييم.

الخوارزمية المقترحة

أُعدّت خوارزمية لحساب قوة الآراء بالاعتماد على القواعد، ويُستخدم العديد من الموارد والأدوات في تطبيق الخوارزمية من أجل إعداد نص المراجعة والمساعدة في مهمة التصنيف المعتمد على القواعد. استخدمنا قوائم الكلمات المقيِّدة للمعنى ومفردات المواقف، حيث إن المعنى الدلالي الإجمالي للجمل يعتمد في المقام الأول والأخير على وحدات الاستقطاب الفردية بها. استخدمنا نسخة مترجمة من مفردات الآراء للإجابة على الأسئلة متعددة الأبعاد (MPQA) في النظام المقترح، كما استعنَّا بكلمات مقيِّدة للمعنى خاصة باللغة العربيّة للمساعدة في تحديد القواعد المطبقة في تحليل النصوص. وتكوّنت مجموعة الكلمات من ثلاث قوائم: أدوات النفي والكلمات المشدِّدة للمعنى والروابط. وتكوّنت هذه الكلمات المقيِّدة للمعنى من كلمات كاملة أو بادئات متصلة بالكلمات.

حساب تصنيف المراجعة

تُمنح كلّ مراجعة نقاطا إيجابية أو سلبية نهائية بناءً على تأويل كل نص من النصوص المحلَّلة باستخدام المنهجية المقترحة. وتُحتسب هذه النقاط بحسب الكلمات الموجودة في مفردات المواقف، على أن تُمنح كلمات المفردات الضعيفة نقطة واحدة، ونقطتين لكلمات المفردات القوية،

إضافة إلى تحديثات النقاط بناءً على الكلمات المقيِّدة للمعنى وفقًا لقواعد النحو.

ويُحتسب التصنيف الكلي للمراجعة، كحاصل قسمة النقاط الإيجابية مقسومة على نقاط الاستقطاب الإجمالية للمراجعة (لكل من النقاط الإيجابية والسلبية)، للحصول على تقدير لنقاط استقطاب النص نسبةً من وحدات الاستقطاب. من ثم يُقارن التصنيف مع التقييم الرقمي الأصلي لمعرفة قوة الاستقطاب للآراء في النصوص، ويقاس عن طريق متوسط الفرق المطلق بين التقييم الأصلي والتصنيف المحتسب.

النتائج والتقويم

يُعيّن الفرق المطلق بين المراجعة الأصلية وتصنيف المراجعة المتنبأ به لكل مراجعة. بلغ الفرق المطلق ١٫٢ للمراجعات الإيجابية (الانحراف المعياري ١٫٣) و١٫٥ للمراجعات السلبية (الانحراف المعياري ١٫٤)، وبالتالي بلغ متوسط الفرق بين تصنيفات الفئتين ٢٫٣ نقطة، مقارنة بالتقييمات التي تضمنتها المراجعات نفسها. وهذه النتيجة تشير إلى مدى مطابقة نقاط رأي المراجعات مع التقييم الأصلي، كما تعكس وجود وحدات استقطابية كثيرة في النصوص المدروسة. وبما أنّ مقاربتنا تُعدّ جديدة التطبيق، فإنّ النتائج تدل على دلائل قوية لإمكانية استخدام الطريقة المعتمدة على القواعد في تصنيف الآراء باللغة العربيّة.

كثيرًا ما يُعد استخدام الطرائق المعتمدة على القواعد في نمذجة اللغة وتحليلها مهمة صعبة، ولا سيما مع اللغات التي تتصف قواعدها الصرفية بالتعقيد مثل اللغة العربيّة. لكن الخوارزمية المقترحة تبيِّن أن استخدام هذه القواعد البسيطة وقوائم الكلمات المخصصة من شأنه تحقيق نتائج واعدة في تصنيف المواقف باللغة العربيّة. وتؤدي الخوارزمية وظيفة تفكيك بنية النحو العربي ونمذجته مع التركيز على وحدات عبارات الرأي، بما فيها الكلمات الاستقطابية وأدوات النفي البسيطة والكلمات المشدِّدة للمعنى والروابط.

إنّ النظام المقترح يستخدم النقاط الاستقطابية المشتقة من الطريقة المعتمدة على القواعد لتحديد التصنيف الكلي لاستقطابية المراجعة، واتضح أنه يقترب كثيرًا من تقييم صاحب

المراجعة. وبالتالي، فإنّ هذه الطريقة الجديدة لاحتساب قوة الآراء تقدم مقاربة قابلة للتوسعة من أجل تحليل المواقف في اللغات المعقدة.

ومن أجل تحسين المقاربة المستخدمة، يمكن توسعة قوائم الكلمات والقواعد وكذلك نمذجتها على نحو أكثر دقة بما يناسب دقائق اللغة العربيّة. كذلك، من المهم استخدام المزيد من المعلومات الدلالية ونمذجة العلاقات بين الكلمات في النصوص، وذلك في البحوث القادمة.

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Pages 165-173

Year of publication: 2014

A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, Wassim El-Hajj

معجم عربي واسع النّطاق للتّنقيب عن العَواطِف (المَشاعِر) في الرأي العَربي

المُقدمة

تَعتمد مُعظم أساليب استخراج الرأي باللغة الإنجليزية بامتياز على معاجم المشاعِر أو العَواطف، مثل مُعجم شبكة (English SentiWordnet) الإنجليزية. وبُذلت العَديد من الجُهود لبناء مَعاجم عربيّة للمشاعر، غير أنها تُعاني من العديد من أوجه القُصور، مثل: محدوديّة الحَجم، وعدم وضوح خُطّة قابلية الاستخدام، نظراً لثراء نظام الصّرف العربي أو بسبب عدم توفره للعامّة.

في هذه الورقة، تمّت مُناقشة كلّ هذه القضايا وتمّ إنتاج أول قاموس عربي واسع النطاق للمشاعر أو العواطف باستخدام مزيج من الموارد المتاحة حالياً، مثل: شبكة English SentiWordnet؛ وشبكة الكلمات العربيّة (Arabic WordNet)؛ والمحلّل الصّرفي للغة العربيّة القياسية(SAMA). وقُمنا بالمقارنة والتّوليف بين طريقتين لبناء هذا المعجم مع إيلاء العناية للهجات العربيّة واللغات الأخرى ذات المَصادر اللغوية القليلة. كما قمنا أيضاً بإجراء تقييم خارجي حول تحليل الذاتية والعواطف أو المشاعر.

وفيما يلي أهم المباحث الرئيسية التي تناولتها الورقة:

أولاً: أساليب (منهجيات) صناعة المُعجم

قمنا بتحديد مُعجمنا العربي للعواطف المستهدف مصدرا يمزج بين العناوين الرئيسية العَربيّة أو الفَرضيات المساعدة التي يشتمل عليها المحلّل الصّرفي SAMA ودرجات تحليل العواطف مثل تلك المستخدمة في شبكة ESWN (إيجابي، سلبي، محايد). وفيما يلي أهم الموارد التي استعنّا بها. وتم لاحقاً استعراض منهجيتين لبناء المعجم المستهدف.

الموارد: تم الاعتماد في بناء المعجم على أربعة موارد متاحة لبناء المعجم وهي: English WordNet، Arabic Word-Net، English SentiWordNet، SAMA. وتمّ إجراء مقارنة بين خصائص الموارد الأربعة مع خصائص المعجم المستهدف من حيث مُحتواها من العواطف والدلالات والعناوين الرئيسيّة.

المنهجية القائمة على الشبكة العربيّة (Arabic WordNet): في هذه المنهجية، اعتُمد على وجود مصدر مشروح غني بالمفردات، وهو شبكة الكلمات WordNet، التي تتماشى مع مُعجم شبكة ESWN. وبالنّسبة للغة العربيّة، فإنّ هذه المنهجية تتطلب خطوتين: المواءمة بين خريطة الشبكة العربيّة AWN وشبكة ESWN، ثم المواءمة بين خريطة SAMA والشبكة العربيّة AWN. تتيح لنا الخريطة الأولى (AWN مع EWSN) الحُصول على درجات العواطف، في حين تتيح لنا الخريطة الثانية (AWN مع SAMA) الحُصول على العناوين الرئيسيّة الصحيحة للكلمات في الشبكة العربيّة AWN. ولذلك نشير إلى المعجم الناتج بـ (ArSenL-AWN).

المَنهجيّة القائمة على قاموس المُصطلحات الإنجليزية (English Gloss-based Approach): في هذه المنهجية، استفدنا من المصطلحات الإنجليزية المرتبطة بإدخالات العناوين الرئيسية لـلمُحلل الصرفي (SAMA)، فقمنا بإيجاد قائمة المترادفات في EWSN بأعلى درجة من التداخل مع المصطلحات الإنجليزية في المحلل الصرفي لكل إدخال.

المَزْج بين المَنهجيّتين: وهنا تمّ عمل توليفة من المعجمين من خلال توحيدهما في مُعجم واحد أطلقنا عليه اسم «ArSenL". ويتألّف المعجَم الموحّد من دمج المصدرين معاً وإضافة حقل جديد في المعجم لتمييز المصدر الأصلي للإدخال.

ثانياً: التقييم

قمنا بإجراء تقييم جوهري لمقارنة الإصدارات المختلفة من معاجم (ArSenL) وتطبيقاتها على مهمة تحليل المشاعر، كما أجرينا أيضاً مقارنة مع أداء مُعجم (SIFAAT).

إعدادات تجريبيّة

أجرينا تجاربنا على نفس المتن الذي استخدمه عبد المجيد وآخرون (٢٠١١)، ويتكوّن هذا المتن من ٤٠٠ وثيقة مأخوذة من (Penn Arabic Treebank – الإصدار ٣) وهي مُجزأة ومُروّسة بطريقة مُمتازة. ووسمت الجُمل حسب الفئات التالية: موضوعيّة، وذاتيّة–إيجابيّة، وذاتية–سلبيّة وذاتية–مُحايدة.

وأُجريَت تجرُبتان لتقييم أثر المعاجم المختلفة على عملية استخراج الرأي، اعتمدت التجربة الأولى أسلوب تحليل الذاتيّة بحيث تُصنّف الجُمل إما إلى ذاتية أو موضوعية. وفي هذه التّجربة، ضبطت المعلومات لتَعظيم قيمة F1 الناتجة للتنبُؤ بجمل الذاتية. أما التجربة الثانية فاعتمدت أسلوب تصنيف المشاعر أو العواطف، بحيث لا تصنّف إلا الجُمل الذاتية إما إلى إيجابيّة أو سَلبية، ويتم تجاهُل الجُمل الذاتية المحايدة. وفي هذه التجربة، تمّ ضبطت المعلومات لتَعظيم متوسط قيمة F1 الناتجة للوسم الموجب والسالب، ثمّ قمنا بتدوين مقاييس الفئات الفردية ومتوسطاتها.

النتائج

أُجريت ثلاثة تقييمات لمقارنة أداء معاجم تحليل المشاعر التي جرى تطويرها:

أولا: قُيّم نطاق تغطية كل مُعجم من المعاجم المختلفة، وعُرّفَت التغطية بنسبة العناوين الرئيسيّة (باستثناء عَلامات الوَقْف) التي تمّت تغطيتها من كُلّ مُعجم. وأظهرت النّتائج بأنّ نِسبة تغطية المعجم (ArSenL-AWN) كانت أقلّ من نِسبة تغطية المعجم (SenL-Eng). في حين كان المعجَم الموحّد هو الأعلى من حيث نسبة التغطية، ويرجع ذلك لوجود عدد أكبر من

العناوين الرئيسية المدرجة في المعجمين الإنجليزي والموحّد. أما من حيث تصنيف الموضُوعيّة، فقد أظهرت معاجم (ArSenL) أداءً أفضلَ مقارنة بخط الأساس الغالب بل وتفوّقت على مُعجم (SIFAAT) من حيث قيمة F١. وبالمجمل، فقد أظهر مُعجم (ArSenL) الموحّد أفضل أداء من بين جميع المعاجِم. وبالمثل، فقد أظهرت تجربة تصنيف المشاعر بأن معاجم (ArSenL) كانت تُعطي باستمرار نتائج أفضل من مُعجم (SIFAAT) ومقارنة بخَط الأساس الغالب. من جهة أخرى، أثبت قاموس (ArSenL) الموحّد أنه يتفوّق على جميع المعاجم الأخرى بجميع المقاييس دون أي استثناء.

وباختصار، يمكن ملاحظة أن المعجم القائم على اللغة الإنجليزية قادر على إعطاء نتائج تفُوّق المعجم القائم على الشّبكة العربيّة (AWN). أما عند الجَمع بين المصدرين، من خلال الدّمج أو التوحيد، فقد سمح ذلك بمزيد من التحسين في أداء تحليل الذاتية والمشاعر. ومن الجدير بالذكر أيضاً أن المعجمين الإنجليزي والموحّد تفوّقا باستمرار على مُعجم (SIFAAT) على الرغم من أنّ الأخير تمّ اشتقاقه يدوياً من نفس المتن الذي استخدمناه للتقييم.

ثالثاً: الخاتمة والتوصيات

قدمت هذه الورقة بتقديم معجماً كبيراً للمشاعر أو العواطف العربيّة باستخدام منهجيات مختلفة مرتبطة بأسلوب شبكة English SentiWordnet. وتم إجراء مُقارنة بين الطّريقتين، حيث أظهرت نتائجُنا أن استخدام الربط القائم على اللغة الإنجليزية يُعطي بالمتوسط أداءً مُتفوّقاً بالمقارنة مع استخدام الأسلوب القائم على شبكة الكلمات (WordNet). في حين أن توحيد المصدرين أفضل من الاعتماد على أيٍّ منهما بشكلٍ منفرد، حتّى إنه يتفوّق على الجودة العالية لاشتقاق الصّفة العاطفية يدوياً من النّص العربي بواسطة المعجم. وتوصي الدّراسة بأن تكون هناك خُطط مُستقبليّة للاستفادة من هذا المعجم لتطوير أنظمة أكثر قوةً لتحليل الذاتية والعواطف. كما أوصت الدراسة أيضاً بوضع خُطط لتوسيع الجهد المبذول ليشمل اللهجات العربيّة ولغات أخرى.

2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)

Year of publication: 2015

Enhancing the Determination of Aspect Categories and Their Polarities in Arabic Reviews Using Lexicon-Based Approaches

Islam Obaidat, Rami Mohawesh, Mahmoud Al-Ayyoub, Mohammad AL-Smadi and Yaser Jararweh

تحسين إيجاد تصنيف المواضيع وقطبياتها في المراجعات العربيّة باستخدام طرق معتمدة على القاموس

إنّ التحليل الوجداني هو عملية تحديد الحس في النص المكتوب باللغات الطبيعية لتحديد إن كان إيجابيا أو سلبيا أو محايدا. يتضمن هذا البحث إعداد مجموعة تدريب لتتم معالجتها لاستخلاص المصطلحات المعجمية، ثم توظّف المعجم المبني لاتخاذ قرار حول كلّ مصطلح، ثم تُقيمّ جودة كل قرار. ومجموعة البيانات المستخدمة هي مجموعة كبيرة من مراجعات الكتب تحتوي أكثر من ٦٣ ألف مراجعة كتاب.

وهناك طريقتان لبناء المعجم: يدوية وأوتوماتيكية، بناء على جذور الكلمات أو ذخيرة لغوية مشروحة، والطريقة الإلكترونية تعاني عادة من قلّة الدقة والمتانة ولكنها تتطلب جهدا إنسانيا بسيطا وتوفر الجهد والوقت كما يسهل تحديثها باستمرار، على العكس من الطريقة اليدوية.

يُحسن بناء المعجم تدريجيا من خلال عملية تحديد قطبية فئة المنظور. وتتألف أول طريقة لبناء المعجم من جدول بخمسة أعمدة، واحد للكلمة أو المقطع الصوتي والباقي للترددات في المراجعات السلبية والإيجابية والمحايدة والمتضاربة، ولتحديد قطبية أيّ مراجعة فإنها تُقسّم إلى كلمات ليتم البحث عن كل كلمة في المعجم لتحديد ترددتها في المعاجم الإيجابية والسلبية

والمحايدة والمتضاربة، ثم تُجمع مرات ظهور الكلمات في المراجعة لإنتاج مصفوفة تمثل الترددات الكلية لكل كلمات المراجعة في كل من المعاجم الأربعة، ويُتخذ القرار بأخذ القطبية المقابلة لأعلى قيمة في المصفوفة، ودقة هذه الطريقة قليلة وبالتالي بحاجة إلى تحسين.

هناك مشكلة تتعلق بالطريقة الأولى وهي أن وجود كلمة واحدة ذات تردد عال باعتبارها كلمة إيجابية في جملة سلبية قد يغير الحكم على الجملة من سلبي إلى إيجابي، ولذا جاءت الطريقة الثانية بحيث يتم عمل أربعة معاجم منفصلة ووضعها في ملفات مختلفة كل منها يمثل القطبيات الأربعة (سلبي، إيجابي، محايد، متضارب) وتوضع كل كلمة في المعجم الذي يكون لها فيه أعلى تكرار وجود، وبالتالي يكون لهذه الطريقة دقة أعلى من الأولى وصلت إلى ٢٩٪.

تحسّن الطريقة الثالثة الطريقتين السابقتين بطريقتين؛ فهي تراعي الأوزان بشكل أدق في المعاجم التي يتم بناؤها، وتستخدم أدوات جاهزة مستخدمة في اللغة الإنجليزية، فإن لم تعثر على الكلمة في المعجم فإنها تترجمها للإنجليزية وتأخذ قطبيتها ثم تضيفها إلى المعجم وقد وصلت دقة هذه الطريقة ٧١٪.

لتحديد فئة المفهوم يُستعاد أقرب جملة في مجموعة الاختبار لكل جملة اختبار، ويُحتسب التشابه بطريقة مشابهة لما سبق، وأول طريقة لبناء معجم المفاهيم مشابهة للطريقة الثانية المذكورة أعلاه والاختلاف الرئيس بينها هو أننا نبني ١٤ معجما لكلّ من فئات المفاهيم الأربع عشرة، ويتم اتخاذ القرار بطريقة مشابهة، أما في الطريقة الثانية فنوظف تقنيات أفضل لمنح أوزان لكلمات المعجم.

تُحدد فئة المفهوم للمراجعة في مجموعة الاختبار بطريقة مشابهة لما نقوم به في الطريقة الأولى من طرق بناء المعجم الثلاث، فيتم البحث عن كلّ كلمة في المعاجم الأربع عشرة، وتضاف أوزانها إلى أوزان الكلمات الأخرى في المراجعة، وذلك يعطينا مصفوفة من ١٤ خانة وتختار الخوارزمية فئة المفهوم التي تكون قيمتها الأعلى في المصفوفة، وإذا لم يتم العثور على الكلمة في المعجم فإن الخوارزمية تأخذ الكلمة التي تسبقها والتي تليها وتأخذ معدلهما.

(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 5, 2014

Year of publication: 2014

Opinion Mining and Analysis for Arabic Language

Mohammed N. Al-Kabi, Amal H. Gigieh, Izzat M. Alsmadi, Heider A. Wahsheh, Mohamad M. Haidar

التنقيب عن وجهات النظر وتحليلها في اللغة العربيّة

تشكّل وسائل الاتصال الاجتماعي مكوّناً رئيسياً للشابكة وتشمل شبكات التواصل الاجتماعي والمنتديات ونقاشات المنتديات والمدّونات الصغرى ...الخ، يولّد مستخدمو وسائل الإعلام الاجتماعي مقداراً ضخماً من النصوص والتعليقات على أساس يومي. وتعكس هذه المراجعات والتعليقات وجهات نظر المستخدمين حول قضايا مختلفة مثل المنتجات والأخبار ووسائل الرفاهية أو الرياضة. ولذلك قد تحتاج مؤسسات مختلفة لتحليل هذه التنقيحات والتعليقات. مثال على ذلك، الشركات يهمها أن تعرف الإيجابيات والسلبيات لمنتجاتها وخدماتها من وجهة نظر الزبائن، كما تريد الحكومات أيضا أن تعرف ميول الناس واتجاهاتهم بخصوص قرارات أو خدمات معنية ...الخ. وبالرغم من أنّ التّحليل اليدوي للمراجعات والتنقيحات النّصية أكثر دقةً من الأساليب والطرق الآلية إلاّ أنّه يحتاج وقتا طويلا وباهظ الثمن وقد يكون غير موضوعي أيضاً. هذا بالإضافة إلى أن الكميات الهائلة من المعلومات الموجودة في شبكات التواصل الاجتماعي يمكن أن تجعل من غير العملي أن يجرى التحليل يدويّاً .

يركزّ هذا البحث على تقييم المحتوى الاجتماعي للغة العربيّة، ويُعدُّ الشرق الأوسط في الوقت الحالي منطقة غنيّة بالتغيرات الاجتماعية والسياسية. وقد يكون الإعلام الاجتماعي مورداً غنياً للمعلومات لتقييم حالات كهذه. في هذا البحث طُوّرت أداة بحث واستخلاص لوجهات النظر، وذلك لجمع أشكال مختلفة من نصوص اللغة العربيّة، سواء أكانت اللغة العربيّة الفصحى أم المحكية. تقبل هذه الأداة وجهات النظر والتعليقات باعتبارها معلومات

وتولدّ مخرجات ونتائج على أساس الاستقطاب المتعلق بالتعليقات.

بالإضافة إلى ذلك، تستطيع هذه الأداة تحديد وجهة النظر أو التعليق سواء أكان موضوعيا أم غير موضوعي، سلبيا أم إيجابيا، قويا أم ضعيفا. وقد أظهر تقييم الأداء لهذه الأداة المتطورة بأنها تعطي نتائج أكثر دقة عندما يتم تطبيقها على أساس التصنيف الموضوعي بالمقارنة مع تطبيقها بشكل عام على النصوص العربيّة دون تصنيف.

قدم هذا البحث أداة أساسية يمكن استعمالها لتحليل المراجعات والتعليقات العربيّة سواء كانت باللغة العربيّة الفصيحة أو العامية. ولغرض تقييم هذه الأداة كان هناك حاجة لبيانات مرجعية لغرض فحص فاعليتها. وبما أنه لم يتم العثور على مثل قاعدة البيانات المرجعية هذه، فقد جمعت كمية من المراجعات العربيّة التي تستعمل اللغة العربيّة الحديثة المستعملة في دول الخليج وبلاد الشام والعراق ومصر. واعتُمدت الأداة المستعملة على المعجم، فقد استحدث معجم من النصوص المجمّعة وذلك بطريقة يدوية. إنّ هذه الأداة قادرة على تمييز الاستقطاب (polarity) و الموضوعية والشدة أو العمق لكل مراجعة عربيّة أو تعليق. كان عدد المعاجم التي بنيت ١٨ معجما. اثنان للمواضيع العامة للتعرف على الاستقطاب و ١٦ معجما بنيت لتشمل ٨ مواضيع مختلفة، هي التقنية والكتب والتعليم والأفلام والأماكن والسياسة والمنتجات والمجتمع.

احتوى الجزء الأخير من هذا البحث على تقييم كفاءة الأداة التي بُنيت، وكانت الدقة في تبويب التعليقات ٩٣,٩٪، والدقة في التعرف على الاستقطاب ٩٠٪، ودقة شدة التعليقات أو عمقها كانت ٩٦,٩٪. وقد أُجريت الفحوصات للتعرف على أسباب الأخطاء التي كان منها الأخطاء الإملائية وقصر بعض التعليقات. ويؤمل توسيع هذه الدراسة مستقبلا باستعمال بيانات أضخم. إن الأداة غير صالحة للتعامل مع الإيحاءات ولغات المحادثة والعربيزي ويؤمل مستقبلا شمول ذلك. كما يؤمل تكوين معجم آلي لكي يستعمل في البحث.

٣-٤-١١ أبحاث التعليم والتعلم الآلي

تضم ثلاثة أبحاث، اثنان منها نوع (أ) هما: أسلوب مُبتكر لتعليم الكتابة العربيّة باستخدام نظام للتعرف على الخَطّ العربي اليدوي، ونمذجة لعبة تعليمية لتلاوة القرآن للأطفال الصم تدعى «mFakih». وبحثان نوع (ب) هما: الاستدلال اللغوي الطبيعي في العربيّة باستخدام نسخة موسعة من شجرة تحرير المسافة مع الفروع، ودراسة حول التعلم العميق لمعالجة اللغات الطبيعية للعربية.

International Journal of Engineering and Advanced Technology Studies, Vol.3, No.7, pp.55-63, September 2015

Year of publication: *2015*

Innovative Approach to Teaching Arabic Writing by Using a Handwriting Recognition System

Abdelkarim Mars

أسلوب مُبتكر لتعليم الكِتابة العَربيّة باستخدام نظام للتعرف على الخَطّ العربي اليدوي

المقدمة

يعرض هذا البحث كيفية تطوير نظام للتعرف على الخَطّ العربي اليدوي أو الكتابة العربيّة على الشّابكة، وهو نظام قائم على أسلوب الشّبكات العَصبيّة. ويُقدّم النظام حُلولاً لمعظم الصعوبات المرتبطة بالتعرف على الكتابة العربيّة أو الخط العربي، كما أن هذا النظام المقترح سيتم دمجه في «أداة تعليم اللغة بمساعدة الحاسوب» لإنشاء أنشطة تعليمية وإيجاد ردود فعل (تغذية راجعة) كافية. وتستخدم هذه البيئة أدوات معالجة اللغة الطبيعية، وبشكل أساسي نظامنا الخاص بالتّعرف على الكتابة اليدويّة، لإنشاء أنشطة تعليمية متعددة، وسيُستخدم هذا النظامُ المطوَّر أداةً لتعليم وتدريس اللغة العربيّة لغةً أجنبيّةً.

مَنهجيّة الدراسة

لأجل حلّ بعض الإشكالات التي تعترض عمل أنظمة تعليم اللغة بمساعدة الحاسوب (Computer Aided Language Learning(CALL)) قُمنا بتطوير نظام لتعليم اللغة يعمل بمساعدة الحاسوب، وهذا النظام مُصمم لدارسي اللغة العربيّة لغةً أجنبيةً أو لغةً ثانيةً باستخدام نظام التّعرف على الكِتابة اليدوية. لكن قبل البدء بوصف خطوات إنجاز هذا النِظام (للتعرف

على الكتابة اليدوية) لابدّ من دراسة خصائص اللغة العربيّة وما تنطوي عليه من تعقيدات وتحدّيات.

الخط العربي (الكتابة العربيّة)

تُعدُّ المعالجة الآلية للغة العربيّة مُهمّةً صَعبة للغاية، خاصةً عند التعامل مع الكتابة أو الخط، وذلك للأسباب التالية:

معظم الحُروف العربيّة يتغيّر شكلها حسب موقعها في الكلمة.

تحتوي العديد من الحروف العربيّة على نقاط وشرطات ملتصقة بالحروف، يضيفها الكاتب على نهاية الكلمات المكتوبة.

في الكتابة العربيّة تتفاوت طُرز الكتابة من كاتب لآخر.

يجب أن تُكتب الكلمة العربيّة بشكل مُقوّس (ملتوٍ)، بحيث تتصل الحُروف بالحروف الأخرى في وسط الكلمة.

أسلوب الشّبكات العَصبية

هناك العديد من طرق التمرن على أنظمة التعرف على الكتابة اليدوية، ونذكر من ضمنها: نموذج ماركوف الخَفي؛ والشّبكات العصبية، ونظام الخبير، وأسلوب الجوار الأقرب (neighbor k-nearesr)، الخ. ويذهب بعضُ الباحثين إلى تقسيم هذه الطُرق إلى فئتين رئيستين:

دلاليّة: وهي الفئة التي تشتمل على وصف أشكال الحُروف بأسلوب مختصر.

إحصائية: حيث يتعلّم النظام بشكل مباشر من البيانات دون الحاجة إلى تخصيص هيكل أو تركيب لنظام المعرفة.

وفيما يخص نظامنا المقترح، فقد اخترنا تبني أسلوب الشّبكات العصبية، وتُعد من الناحية النموذجية مزيجاً من الوظائف الأولية التي تُسمّى الخَلايا العَصبية أو (النيورونات)، بحيث

تُدرك كل خلية عصبيّة اصطناعية في كل مرة وظيفةً غير خَطيّة، وتمثل في الشبكة مخرجات الخلايا العَصبيّة (النيورونات).

وتتألّف الشبكة العصبيّة من جُزأين، الأول يُسمّى الإدراك مُتعدد الطبقات (MLP)، والثاني يُسمّى (Time Delay Neural Networks) ويرمز لها بالرمز (TDNN)، وتحتوي على ثلاث طبقات: طبقة إدخال (مُدخلات)، وطبقة واحدة أو أكثر مَخفيّة، وطبقة إخراج (مُخرجات). وتوصل الخلايا العصبية في كل طبقة بجميع الخلايا العصبية الأخرى من الطبقة السفلى، وفي هذه الحالة، تصبح الشبكة مُتّصلة بالكامل. أما في مجال التعرف على الكتابة اليدوية، فقد استخدمنا شبكة (TDNN) للتحرك المكاني، وهي شبكة غالبا ما تستخدم في الوقت المناسب للحصول على بيانات متتابعة، لذا من المناسب تماماً استخدامها لتمييز خط اليد.

عملية التعرف أو التّمييز

إنّ نظام التعرُف على الكِتابة اليدوية هو تقنية مبتكرة لمعالجة اللغة الطبيعية من خلال التعامل مع الحِبر أو المِداد مُدخلا بواسطة جِهاز مُتصل بالشّابكة (حاسوب لَوحي أو أقْلام رقَميّة). ولتحويل هذا الحِبر أو المِداد إلى نصّ رقمي، يقوم النظام بواسطة أداة خاصة بإجراء عدّة مُعالجات، من ضمنها المعالجة المسبقة والتّطبيع. وبعد المعالجة المسبقة للحبر تقوم أداة أخرى باستخراج الخصائص الضرورية التي ستستخدم بعد استخراجها في أداة التصنيف اعتماداً على الشبكة العصبيّة لإعطاء ماهية الكلمة.

وقد اشتملت خوارزمية نموذجنا المقترح للتعرف على أربع خطوات:

المعالجة المسبقة والتطبيع.

التقسيم أو التجزئة.

استخراج السِمة.

التمرن (التدريب) والتعرّف.

تقييم عملية التعرُف

أجرينا تقييما لقاعدة بيانات الاختبار مستخدمين مُعدّل التعرف الصّحيح لتقييم أداء نظامنا. وقد أظهرت نتائج تقييم النظام باستخدام قواعد بيانات الاختبار أنّ نتائجنا كانت ممتازة، حيث أمكنَ التعرف على مُعظم الحُروف.

تكامل أو إدماج نظام التعرُف على الكتابة اليدوية

سنقوم لاحقاً بإدماج نظامنا للتعرف على الكتابة العربيّة من خلال تكييف احتياجات مُخرجاته من اللغويين. وبشكل عام، يجب على المتعلّم، أثناء عملية تعلّم الكتابة، البدء بكتابة الحُروف، ثم الكلمات، وأخيراً كِتابة الجُمل. لذا، قمنا بدمج نظامنا هذا مع ثلاثة نماذج للتعرف، وهي: نموذج الحُروف، ونموذج الكلمات ونموذج الجُمل. وتتكوّن عملية إيجاد النشاط من ثلاث خُطوات، أولاً: يجب على المعلّم وضع سياق تعليمي لتطبيق الخط (الكتابة)، ثانياً: على المعلّم اختيار نموذج التعرف الذي سيُستخدم، وأخيراً: يجب على المعلّم وضع التعليمات والتحقق من صِحّة تكوين النشاط.

النتائج والمناقشة

في نهاية مرحلة تنفيذ البناء الكامل للنظام، أغفلنا فقط – خطوة اختبار النظام وتقييمه، غير أنّ تقييم بناء النظام ومساهمته في جودة تعليم اللغة العربيّة وتعليم اللغة بمساعدة الحاسوب يُعدّ مشكلة. كما أنّ تطوير أداة معالجة اللغة الطبيعية وإدماجها في نظام تعليم اللغة مُكْلف جداً من حيث الموارد والتنفيذ والوقت والجهد.

ولغرض اختبار المكونات المختلفة للمنصة قبل وضعها ضمن بيئة تعليمية حقيقية، قمنا بإجراء بعض الاختبارات التجريبية على بعض المعلّمين والدارسين، فمكنتنا هذه الاختبارات من الوقوف على مزايا نظامنا بالنسبة للمُعلّمين والدّارسين من جوانب عدة. وبعد استكمال الاختبارات على نظامنا في الجامعات الألبانية والمولدافية، خططنا لتطبيق بيئة النظام على الشّبكة وتوفيرها لمجتمع مُعالجة اللغة الطبيعية ومجتمع تعليم اللغة بمساعدة الحاسوب.

الخاتمة

هناك حالياً فرصة واعدة في الأسواق للباحثين عن أسلوب ذكي لتعليم العربيّة، لأنّ مجال تعليم اللغة العربيّة بمساعدة الحاسوب ما زال يقتصر على بعض النماذج الأولية. وتُعدُّ هذه الدراسة أن النظام المقترح لديها منفتحٌ لتحقيق نقلة نوعية نحو تطوير هذه المَنصّة في المستقبل. وفي هذا الصّدد، يمكننا أن نضيف للنظام عددا لا نهائيا من الأنشطة من خلال دمج العديد من أدوات مُعالجة اللغة الطبيعية أو البرمجة اللغوية العَصبيّة مع العديد من الموارد، وبحيث تكيّف المنصّة مع احتياجات المعلّمين والتربويين.

International Journal on Islamic Applications in Computer Science And Technology, Vol. 2, Issue 2, June 2014, 8-15

Year of publication: 2014

mFakih: Modelling Mobile Learning Game to Recite Quran for deaf Children

Azham Hussain, Nazean Jomhari, Fazillah Mohmad Kamal, Normala Mohamad

نمذجة لعبة تعليمية لتلاوة القرآن للأطفال الصُم تدعى "mFakih"

المقدمة

تعليم القرآن للطلبة الصُم جاء متأخراً أكثر بكثير بالمقارنة مع تعليمه للطّلبة المكفوفين؛ فالصُم غير قادرين على السّماع، ويواجه المعلّمون نتيجة لذلك صعوبةً في تعليمهم. إن برنامج (mFakih) هو أسلوب مُبتكر لتعليم اللغة العربيّة للأطفال الذين يعانون من ضَعف السَمع، حيث تُساعد هذه التّقنية المربّين على تعليم الطّلبة الصُمّ تلاوة القرآن بطريقة أكثر عملية باستخدام الأعداد والألوان. مع ذلك، هناك بعض الأطفال الذين عانوا من المتاعب نتيجةً لاضطرارهم للانتباه لنفس الكيان (entity) كلّ يوم. من هنا، صَمّمت هذه الدراسة نموذجاً يتكوّن من المتطلبات (المحتوى) والبنية الهندسية للعبة تعليمية على الهاتف المحمول، يمكن استخدامها لتطوير تطبيق على الأجهزة المحمولة. وتحققت الدراسة من صحة النموذج من خلال تطوير نموذج أوّلي، فأظهرت النتائج أن النموذج الأولي يمكن تطويره باستخدام النموذج الذي أُنشِئ وأُطلق عليه اسم (م-فقيه mFakih).

مُشكلة البَحث والمُبررات

تشهد التكنولوجيا تطوراً جديداً كلّ يوم، ومع هذه التطورات بدأنا نلمس بعض التّحسينات المذهلة التي يشهدها مُجتمع الصُمّ. فلُغة الإشارة باتت تُمثّل وسيلة رائعة للتواصل

ما بين الصُمّ، بعدما كانت هذه اللغة تمثّل جُزءاً من وسائل الاتصال بينهم عبر آلاف السّنين. لقد كانت هذه اللغة الشكل الوحيد المتاح للاتصال بالنسبة للكثير من الصُمّ. وتمثل «رموز الإشارة الهجائية» لغة إشارة من الحروف العربيّة، التي تستخدم لُغةً للإشارة للتواصل بين الطلبة الصُمّ.

إن ضَعف التفاعل في الفُصول الدراسية التّقليدية يُسبّب عادةً عُزوفاً لدى الطلبة عن عملية التّعلّم في مُنتصف الطّريق. علاوة على ذلك، هناك الكثير من التّطبيقات المتقدمة لتعليم لغة الإشارة عبر الهواتف النقالة، لكن تفتقد أكثر هذه التّطبيقات للجودة، حيث إن جَودتها أقل بكثير من أن تُمكّن الناس من فَهم أساسيّات برنامج فقيه وأسلوبها بشكل جيّد. ويمكن للطّلبة الصُمّ تحسين تلاوتِهم للقرآن الكريم، كما يمكنهم الاستمتاع بقراءة القرآن الكريم من خلال مُمارستهم لهذه اللعبة.

منهجية الدراسة

قُسّمت منهجية هذه الدراسة إلى أربع خطوات رئيسية، الخطوة الأولى تمثلت في إجراء دراسة مستفيضة للجوانب المختلفة للتعليم بواسطة الهاتف المحمول والألعاب التعليمية، وهذا يشمل تحديد الميزات الجديدة للتعلّم عبر ألعاب الهاتف المحمول، والاتجاهات السابقة والحالية له. وللحصول على أساس جيد من المعرفة وفهم المتطلبات، وأدوات تطويره ومكامن القوة والضعف في الألعاب التعليمية عبر الهاتف المحمول؛ أما الخطوة الثانية فتمثلت في التحليل، وركزت على كيفية جمع المتطلبات، فتمثلت أساليب جمع المتطلبات بالمقابلات الشخصية والمشاهدات. لكن كان هناك بعض المتطلبات التي أمكن توفيرها بواسطة مُدرّب فقيه (**Fakih instructor**)، وذلك اعتماداً على الخِبرة والتّجربة في استخدام أسلوب فقيه مع الطلبة الصُمّ. أما الخطوة الثالثة، فهي تصميم النموذج وتطويره لوضع تصوُّر للمتطلبات الناتجة. وبما أن هذه الدراسة تهدف لتطوير نموذج مَرجعي من برنامج فقيه **Fakih**، فقد كانت لغة النّمذجة المستخدمة هي لُغة النّمذجة الموحدة (**Unified Modeling Language**)، وهي أسهل لُغات

النّمذجة المستخدمة بشكل عام في تطوير أنظمة المعلومات؛ وأخيراً تمثلت الخُطوة الرابعة في التقييم، حيث استعانت الدراسة باثنتين من تقنيات التقييم، وهما: مُراجعة الخبير والاستعراض من خلال تطوير نموذج أولي من برنامج فقيه مشابه لـنموذج (شيروود و روت (١٩٩٨)). والغرض من استخدام أسلوب التقييم هو ضمان نجاح التنفيذ النهائي للنموذج المرجعي لفقيه بحيث يُمثّل تطويراً يكون له فوائد مُؤكّدة وفَعّالة من حيث الإنتاجية وجودة المنتج.

نَمذَجة (م- فقيه mFakih)

ناقشت الورقة في هذا المبحث متطلبات النظام بما في ذلك المتطلبات الوظيفية والاحتياجات غير الوظيفية، فضلاً عن متطلبات النظام التي تغطي احتياجات البرامج والأجهزة وأهداف المستخدِم و الجِهاز المستخدم. كما ناقشت عدداً من تقنيات جمع متطلبات ودراسة نطاق النظام. ونظراً لحجم النّطاق (رموز الإشارة الهجائية العربيّة)، كانت موارد جمع المعلومات المطلوبة محدودة. أما التقنيات التي استُخدمت لجمع المعلومات والمتطلبات لتطوير برنامج فقيه فهي: العصف الذهني، والاختبار الميداني، والبحث على الشّابكة والبحث في الأدبيات المشابهة.

تنفيذ النّموذج الأوّلي

الأداة المستخدمة لتطوير النّموذج الأوّلي هي «إكليبسEclipse"؛ و «إكليبس» هي بيئة لتطوير البرمجيات مُتعدّدة اللغات التي تحتوي على قاعدة أساسية للعمل وبرنامج مُساعد قابل للتوسع في المكونات لتخصيص بيئة عمل مُلائمة. وتتضمن Eclipse SDK مِنصّة أساسية بالإضافة إلى اثنتين من الأدوات الرّئيسية المفيدة لتطوير المكونات الدّاخلية. في حين تنفّذ أدوات تطوير جافا (JDT) بيئة تطوير جافا كاملة المواصفات. من جهة أخرى، تضيف بيئة المطور المساعد (PDE) أدوات مُتخصصة لتبسيط تطوير البرامج المساعدة والمكونات الإضافية الأخرى.

وتُعدُّ مرحلة الاختبار مُهمّةً للغاية خلال دورة حياة تطوير البَرمجيّات، حيث تنتهي كلّ

مرحلة بالتحقُق والتقييم لتحديد جميع الأخطاء في كل مرحلة من المراحل والتخلُّص منها. وعادة تقع بعض الأخطاء في النظام خلال المراحل المختلفة لتشغيله. لذا، ينبغي إجراء الاختبار على مُستويات مُختلفة لتحديد تلك الأخطاء التي حدثت خلال كل مرحلة. وقامت هذه الدراسة باختبار النموذج الأوّلي باستخدام نموذج حالة الاختبار لكل مُستوى. وأظهرت النتائج أن النموذج الأوّلي يعمل بشكل جيّد دون أي أخطاء جَوهرية. على الرغم من ذلك، تحتاج واجهات النموذج إلى تحسين من أجل جذب المستخدمين لتجرُبة اللعبة.

الاستنتاجات

من المتوقع أن تصبح الألعاب مواد تدريسية للمُعلّمين، ومن المتوقع أن تجتذب مشاركة أكبر من الأطفال نظراً للميّزات التفاعلية التي تتمتع بها. كما أنها تُوفّر تغذية راجعة حول تفاعلات المستخدِم، وهذه التّغذية الرّاجعة وتقنيات التفاعل تجعل أسلوب الطالب في التّعامل مع المحتوى أكثر مُتعةً.

لقد طُوِّر النّموذج الأوّلي باستخدام أدوات (**Eclipse IDE**، و **Android SDK**)، ولغتي **Java** و **XML**. وهو تطبيق يعمل على نظام أندرويد للهواتِف الذكية أو الحاسبات اللّوحية. يُزوّد الطالب بهذا التطبيق مجاناً أو بمُقابل، حيث يُساعد جميع الطلبة على تعلُّم اللغة العربيّة، أو تحسين قدرات الطلاب في تعلُّم اللغة العربيّة. ويشمل هذا التطبيق العديد من الوحدات التفاعلية. ويمكن إضافة المزيد من التصاميم أو الوحدات التفاعلية استناداً إلى طبيعة استخدام هذا التطبيق. كما يحتوي هذا التّطبيق على أمثلة وتمارين تستخدم كلمات بسيطة من القرآن الكريم. وفي المستقبل، يمكن تضمين كلمات أكثر تعقيداً أو وحدات تفاعلية مع قراءة القرآن أكثر تنوعاً، وحسب احتياجات المستخدمين، على سبيل المثال من خلال الدمج مع برنامج المترجم (الإنجليزية إلى العربيّة / العربيّة إلى الإنجليزية)، والقراءة التلقائية للنص وغيرها. كما أنه يمكن أيضا تشغيل التطبيق على الهواتف النقالة التي تعمل بأنظمة تشغيل أخرى مثل (**iOS** أو **Blackberry OS**).

Journal of Artificial Intelligence Research 48 (2013) 1-22

Year of publication: 2013

Natural Language Inference for Arabic Using Extended Tree Edit Distance with Subtrees

Maytham Alabbas, Allan Rhk amsay

الاستدلال اللغوي الطبيعي في العربيّة باستخدام نسخة موسّعة من شجرة تحرير المسافة مع الفروع

هدف الباحثون في بحثهم هذا إلى تطوير نظام لمعالجة اللغات الطبيعية (Natural Language Processing (NLP)) في اللغة العربيّة من شأنه تحديد مدى الترابط بين أجزاء النص الواحد، فهناك العديد من التطبيقات التي تعتمد على هذا النظام تتطلب حساب درجة التشابه بين النصوص من الناحية النحوية والدلالية. اعتمد الباحثون على نهج TED+ST الذي يساعد في عملية التحقق من تتابع الأجزاء في النص الواحد. وقد جرى الحصول على نتائج أولية مشجعة مقارنة مع غيرها من الأنظمة، بالإضافة إلى استخدام نسخة موسعة من شجرة تحرير المسافة (Tree Edit Distance) مع TED + ST في مجموعة الاختبار الإنجليزية (RTE-٢). وقد استُخدم هذا النظام في عدد كبير من اللغات الأخرى مثل اللغة الإنجليزية وغيرها بشكل موسع، لكن للأسف، لم تحظ اللغة العربيّة بكثير من الاهتمام في هذا المجال نظرا لعدة عوامل تختص بطبيعتها مثل صعوبة اللغة العربيّة من ناحية الصرف، كما أنّ عدم استخدام حركات التشكيل زاد المشكلة تعقيدا وغموضا، نظرا لاختلاف إملائها واحتوائها على أكثر من معنى، مما أدى إلى ظهور مشكلة في التحليل الصرفي للكلمات بسبب تعدد أشكال الكلمة الواحدة مع اختلاف حركاتها التشكيلية.

ومن التحديات الأخرى التي واجهها الباحثون عدم توفر مجموعة البيانات والموارد الكافية في اللغة العربيّة، التي يمكن الاعتماد عليها من أجل تدريب نظام (TE) في اللغة العربية

وفحصه. لذلك، كان من الضروري اللجوء إلى تطوير بعض مجموعات البيانات عن طريق إجراء بعض التغييرات في مجموعة من الأزواج في الفرضيات الجاهزة في مهام مجموعة الاختبار RTE مثل نصوص وكالات الأنباء، التي تتصف باحتوائها على جملة أو اثنتين وبأن الجمل عادة طويلة نسبيا. لقد أمكن الحصول على مجموعة البيانات العربيّة لأزواج p-h لمهام TE، عن طريق استخدام التقنيات شبه التلقائية، التي مرت بمرحلتين أساسيتين؛ المرحلة الأولى هي المسؤولة عن جمع أزواجp-h بشكل تلقائي من مواقع الويب الإخبارية. فالفكرة الرئيسية هنا أنه تُطرح الاستفسارات لمحرك البحث وتُصفى الردود بشكل تلقائي أيضا. وفي المرحلة الثانية تُشرح هذه الأزواج بشكل يدوي لاتخاذ القرار سواء بإدخالها أو عدمه. وفي هذه المرحلة تخضع البيانات للشرح والتوضيح بواسطة ثمانية خبراء في هذا المجال وثمانية أشخاص عاديين (من غير الخبراء) لتحديد مختلف أنواع الأزواج في المجموعات.

ومن خلال التجارب التي قام بها الباحثون، ثبتت متانة نهجهم المقترح مع وجود محلل أكثر دقة، مما حسّن أداء الأداتينZS-TED و ED+ST ، وساهم في تحسين أداء التقنية العربيّة بما يقارب ٥ ٪ في مقياسF-score و ٤٪ في مستوى الدقة مقارنة مع غيرها من التقنيات المعروفة.

Journal of Computational Science Volume 26, May 2018, Pages 522-531

Year of Publication: 2018

Deep learning for Arabic NLP: A survey

Al-Ayyoub, Mahmoud & Nuseir, Aya & Alsmearat, Kholoud & Jararweh, .Yaser & Gupta, B B

دراسة حول التعلم العميق لمعالجة اللغات الطبيعية للعربية

أحدثت التطورات الحديثة في التعلم العميق (Deep learning) اختراقات في العديد من المجالات مثل معالجة اللغات الطبيعية (NLP) ومعالجة الكلام. وقد تبين أن العديد من مناهج DL تسفر عن نتائج متطورة حول مهام متنوعة ذات أهمية كبيرة للشبكات الاجتماعية على الإنترنت (OSN) والحوسبة الاجتماعية مثل تحليل المشاعر (SA).

تقدم هذه الورقة البحثية مسحًا للأوراق المنشورة حول استخدام تقنيات التعلم العميق في معالجة اللغات الطبيعية، حيث ركز الباحثون على اللغة العربية بسبب أهميتها وندرة الموارد والتحديات المرتبطة بالعمل عليها، حيث لاحظ الباحثون من خلال دراستهم أن تقنيات التعلم العميق لم تحصل بعد على الاهتمام الذي تستحقه من مجتمع البرمجة اللغوية العصبية العربية (ANLP) مقارنة مع الاهتمام الذي تحصل عليه لغات أخرى على الرغم من التبني الواسع لشبكات التواصل الاجتماعي في العالم العربي. تركز معظم الأعمال المنشورة على استخدام التعلم العميق في حل المشاكل المتعلقة بالتعرف الضوئي على الحروف في حين أن الأكثر حداثة وتنوعًا هي تطبيق التعلم العميق في تحليل المشاعر، والترجمة الآلية، وتقطيع النصوص، وما إلى ذلك.

١- التعرف الضوئي على الحروف: واحدة من أبرز اهتمامات التعلم العميق وتشمل عددا من التطبيقات من أبرزها التعامل مع النصوص العربية المكتوبة بخط اليد حيث تُعدُّ واحدة من أكثر مشاكل التعرف الضوئي على الحروف التي يتم تناولها بشكل متكرر، والتي يتم التعرف عليها من خلال النص المكتوب باللغة العربية (Handwritten texts). كذلك

تُعدُّ النصوص المطبوعة من المشكلات التي عولجت باستخدام تقنية التعرف الضوئي والنصوص المرافقة لمقاطع الفيديو (Video-Overlaid Text) طرحت في عدة دراسات. إضافة إلى قضية نصوص المشاهد الطبيعية (natural scene images)

٢- إنشاء التوصيفات للصور (Caption Generation) : تُعدُّ مشكلة إنشاء وصف لصورة معينة تلقائيًا فكرة شيقة وعملية. ولحل هذه المشكلة قامت بعض الدراسات بتقسيم الصورة لأجزاء ومعالجة كل جزء على حدة.

٣- أنظمة التعرف على الصوت: في إحدى الدراسات الأولى التي تستخدم تقنيات التعلم DL للتعرف على الأرقام العربية المنطوقة.

٤- نمذجة اللغة (Language Modeling).

٥- الترجمة الآلية التلقائية (Automatic Machine Translation) هي عملية تطوير أنظمة ATM المستندة إلى تقنية التعلم العميق لترجمة النص من/ إلى اللغة العربية.

٦- اكتشاف اللهجات (Dialect detection): حيث تشتمل توظيف التعلم العميق في التمييز بين اللهجات في اللغة العربية.

٧- التقسيم التلقائي للنصوص (Dialectal Arabic (DA) Segmentation): يمكن أن تكون تجزئة النص المكتوب في DA مهمة صعبة بسبب نقص القواعد والمعايير والموارد.

٨- تصنيف النصوص العربية (Text Categorization): حيث اهتمت تقنيات التعلم العميق بالتصنيف الآلي للنصوص العربية وفقا لموضوعاتها.

٩- تحليل المشاعر (Sentiment Analysis): حيث تعتمد هذه الأنظمة على التعرف الآلي على المشاعر وميول الأشخاص من خلال النصوص المكتوبة.

١٠- أنظمة الأسئلة/ الإجابات (Question Answering): مع تضخم المحتوى المتاح على شبكة الإنترنت، أصبحت أنظمة الأسئلة والإجابة ((Q /A)Question/Answering) من أكثر أدوات البحث شهرةً حيث لاقت اهتمام عدد كبير من الباحثين والمستخدمين. ففي الآونة الأخيرة طُوِّر العديد من هذه الأنظمة وخاصة فيما يتعلق باللغة العربية.

الفصل الرابع
المناقشة والتوصيات

٤-١ الأبحاث العربية

٤-٢ الأبحاث المترجمة

٤-٣ التوصيات

٤-١ الأبحاث العربية:

إن استعراض الدراسات العربية في حوسبة اللغة يظهر أن اهتمامات الباحثين العرب من لسانيين وحاسوبيين في مجال حوسبة اللغة العربية قد بدأ بداية جادة منذ ثمانينيات القرن الماضي، وهي المرحلة التي شهدت دخول الحاسوب في مجالات الاختصاص العلمي وانتشاره ثقافيًا على المستوى العالمي، أما المرحلة التي سبقت ذلك فقد اعتمدت على الإحصاء للمفردات اللغوية وجذورها، وهي أبسط عمليات الحوسبة. ويمكن من متابعة جلّ هذه الدراسات ملاحظة عدد من النتائج التي تظهر واقع الدراسات العربية في حوسبة اللغة، ولعل أهمها:

١- أشار جلّ الدارسين إلى مفهوم اللسانيات الحاسوبية، وميز بعضهم علم اللغة الحاسوبي عن علم الحاسوب اللغوي، ولعل هذا التمييز لم يأتِ على أساس اختلاف الباحثين وطريقة المعالجة فحسب، وإنما بالارتباط بمجال العمل أيضًا، فاهتمام اللسانيين انصب على التصورات المنطقية التي تتجاوز الوصف إلى التوصيف باعتبار ذلك ممهدًا لتهيئة أنظمة اللغة العربية لتتواءم مع الحوسبة، واهتمام الحاسوبيين انصب على المعالجة الآلية القائمة على الخوارزميات.

٢- إن عددًا من الدراسات بما فيها المتأخرة - قد ذيلت عنوانها بإشكاليات وحلول؛ مما يشعر بأن البحوث في حوسبة اللغة العربية ما زالت لم تستقر على ثوابت واضحة على صعيد المصطلحات والمفاهيم وطرق المعالجة، ولعل ذلك راجع إلى حالة التأخر المتابعة والملازمة لانتقال النظريات والمفاهيم والمناهج الغربية، وهو شأن عام، يجعلها غير مستوعبة، ويدفع الدارسين - في الآن ذاته- إلى مواكبة الجديد المتطور.

٣- إنّ من أهم الإشكاليات التي تواجه دراسات حوسبة اللغة عدم وجود نظرية لسانية شاملة في الدراسات اللغوية قادرة على ضم أطراف الظاهرة اللغوية وعناصرها بقدر من الشمول يسمح باختزال أنساق اللغة إلى عدد محدد من الأنساق الرياضية الحسابية، ومن ثمّ بات الاعتماد على الحاسوبيين في تقديم الجانب التقني لحل إشكاليات الجانب المعرفي في النظرية اللسانية، والملحوظ أن جهود الحاسوبيين تفوق ما قدمه اللسانيون في هذا المجال، وهو ما

يجعل سير العمل معكوسًا؛ إذ تسبق الأداة الظاهرة والمادة، وتصبح المسألة وكأن دراسات حوسبة اللغة تسعى إلى تطويع اللغة لتناسب الأداة بخلاف الأصل في هذا الشأن.

٤- إن أغلب اللسانيين لم يستطع أن يضع الأصول الضرورية لحوسبة اللغة؛ إذ بقيت تصوراتهم بمنأى عن فهم العمليات الرياضية التي يتعامل معها الحاسوب، وقد جاءت مجمل تصوراتهم ذات طابع عام أو تقليدي غير ممنهج بدقة علمية كافية، ولعل من أبرز المفاهيم التي ارتكزت عليها بعض الدراسات التركيز على الفرق بين الوصف والتوصيف باعتبار الأخير يقدم بديلاً للحدس عند الإنسان وهي الفكرة التي طرحها نبيل علي وتبناها نهاد الموسى، وعلى الرغم من أهمية هذه الفكرة لتأسيس حوسبة اللغة؛ على الصعيد الاصطلاحي التأسيسي على الأقل، إلا أنها بقية في الإطار العام دون أن توضح أساسيات أصول النظام اللغوي للغة العربية بشكل دقيق، ولذلك جاءت كثير من الدراسات حالمة أكثر منها واقعية عملية.

٥- في المقابل فإن جهود الحاسوبيين في الأغلب - انحسرت في معالجة مستويات دون أخرى مثل المستوى الصرفي، أو مباحث محددة هي التي يمكن إخضاعها للقياس مثل: الجملة البسيطة، والمركب العددي، وهي أسهل مجالات نظام اللغة العربية في المستوى التركيبي النحوي.

٦- إذا كانت بعض النظريات قد أتاحت المجال إلى الضبط المعياري الدقيق للنظام اللغوي مهما اختلفت تمثلات النسق فيه؛ مثل: الميزان الصرفي، والصيغ الصرفية، ونظرية العامل، وتحديد المعاني النحوية وما يرتبط بها من علامات صوتية إعرابية، فإن ما بقي مستعصيًا على الضبط والتقنين السياقات اللغوية التي تعين الرتبة للمفردات اللغوية، وتحدد الارتباطات الدلالية التداولية لها، ثم تتحدد الدلالة وفقًا لذلك كله، وهو ما حاولت دراسات الحوسبة تأصيله وبحثه في دراسات (الأنطولوجيا).

٧- اتسم منهج الحاسوبيين باعتماد المنطق الرياضي المحض، وقياسه تجريبيًا، وهو ما ميز جهودهم بالعلمية والعملية في آن، في حين بقيت جهود اللسانيين أسيرة التصورات ٨-

القواعدية للغة الطبيعية المرتكزة على تصنيفات النحو التقليدي وتقعيده، والغريب أنها عوملت كما لو كانت خوارزمية تمامًا مع أنها مبنية على الحدس الإنساني. ويظهر هذا الفرق مثلاً من أنّ اعتماد الحاسوبيين على الذخيرة اللغوية كان يعني الاعتماد على حالة منظمة من الاستخدامات اللغوية، ومن ثم فدراساتهم تقوم على الواقعي والمتحقق، في المقابل كانت دراسات اللسانيين افتراضية، وتعتمد على الأمثلة المصنوعة، وتوقعات الخطأ والصواب، ومن ثم اتسمت بالعمومية وعدم إدراك طبيعة المجال البيني معرفيًا ومنهجيًا.

٩- ثمة اختلافات واضحة في اهتمامات الحاسوبيين واهتمامات اللسانيين؛ إذ تحتل صنعة المعجم والصرف والترجمة أولوية للحاسوبيين، يلحظ اهتمام اللسانيين بقضايا النحو اهتمامًا ملحوظًا.

١٠- ثمة نظريات لغوية ولسانية قديمة أو حديثة في التعامل مع الظاهرة اللغوية تؤطر بعض المستويات اللغوية، مثل: نظرية الحقول الدلالية لم يطّلع عليها الحاسوبيون، في الوقت الذي اهتموا بهذا المستوى في مجال (الأنطولوجيا)، مما يشير إلى وجود فجوة بين جهود الطرفين، وذلك ما يكشف عن ضرورة التفكير جديًا بالاهتمام بمجال معرفي بيني متخصص باستحداث برامج أكاديمية متخصصة فيه، لا مجرد الاعتماد على جمع جهود الطرفين والتقريب بين خبراتهما.

١١- على الرغم من عناية بعض المؤسسات العربية بحوسبة اللغة العربية، مثل: مجمع اللغة العربية الأردني الذي عقد أكثر من ندوة في مواسمه الثقافية للبحث في هذا المجال، والمنظمة العربية للتربية والثقافة والعلوم، ومعهد التعريب والترجمة في المغرب، ومركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية في السعودية، وغيرها، إلا أنها لم تتجاوز حدود المؤتمرات والندوات إلى تبني مشاريع طويلة الأمد في حوسبة اللغة العربية ببناء وحدات وأقسام ثابتة، تهتم بإنجاز برامج تطبيقية في مجالات المعالجة الحاسوبية للغة العربية، وتكون قادرة على التنبؤ بالحاجات المستقبلية لا أن تهتم بما يطرأ لحظيًا.

٤-٢ الأبحاث المترجمة

٤-٢-١ معايير اختيار الأوراق البحثية

أولاً: راعت اللجنة عند اختيارها الأبحاث التي ترجمت اختلاف اهتمام الناظر إلى موضوع حوسبة اللغة باختلاف خلفيته. فاللغوي يوجه اهتمامه نحو معالجة مستويات اللغة المختلفة من خلال لسانيات الحاسوب. ومن أبرز الأمثلة على هذا النوع البحوث في نظريات تشومسكي في النحو التوليدي والتحويلي. وتشمل بحوث هذا الاتجاه مواضيع الصرف والنحو والدلالة والمعجم.

أما الحاسوبي فيهتم بخوارزميات التحليل والتركيب وتطبيقاتها في معالجة اللغات الطبيعية. ومن أمثلة هذه الخوارزميات آلية فحص الحالات المحدودة التي طبقت في تحليل الصرف.

وقد رأت اللجنة توجيه عملها إلى الجمهور الذي سيتلقاه من خارج بيئة البحث والتطوير التقني، لذلك احتوت البحوث المختارة مقدمات ومفاهيم أساسية في حوسبة اللغة.

ثانيا: راعت اللجنة المراحل التاريخية التي مرت بها بحوث حوسبة اللغة وخطوط البحث الرئيسية التي ظهرت خلالها وأهمها:

البحث المستند إلى حوسبة بنية اللغة وقواعدها وقد بدأت به أبحاث حوسبة اللغة تاريخيا وساد حتى ثمانينات القرن الماضي.

البحث الإحصائي الذي يبني نتائجه من التحليل الإحصائي لكميات كبيرة من النصوص للوصول إلى نتائج حول اللغة دون الخوض في بنيتها وتركيب مستوياتها، وقد ساد منذ فترة التسعينات من القرن الماضي.

البحث الهجين الذي يجمع بين الخطين السابقين وهو أحدث الاتجاهات ظهوراً، وهو ما تبنى عليه حاليا تطبيقات الذكاء الاصطناعي وأنظمة الآلة المتعلمة.

ثالثاً: إن حوسبة اللغة هي تقنية توظف علوما وتقنيات مختلفة، لكنها ليست علما مجردا بذاتها، وهي لذلك سريعة التطور والتغير، وبناء على هذه الحقيقة :

فإن أكثر البحوث القديمة وخاصة في خوارزميات التحليل والتركيب والتطبيقات فقدت قيمتها بظهور خوارزميات وتطبيقات أحدث منها.

وحتى البحوث الجديدة فإنها عرضة لأن تصبح غير ذات قيمة في أي وقت بظهور بحوث أحدث منها.

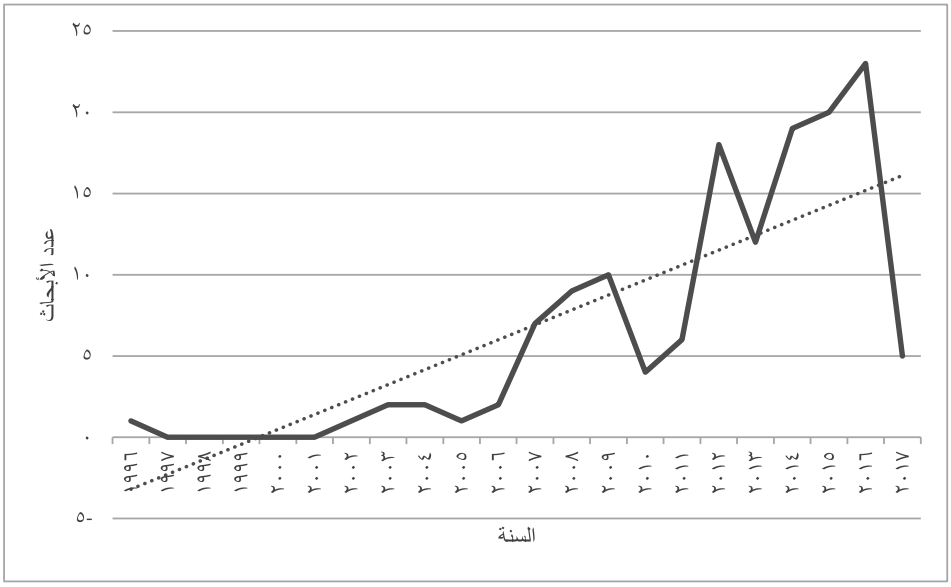
لكن هناك بحوثا يمكن اعتبارها بحوثا مؤسِسة ساهمت في تكوين تيار بحثي أو اتجاهات ويستدل عليها من خلال كثرة العزو لها في البيئة العلمية.

٤-٢-٢ توزيع الأبحاث

وفي ما يأتي جدول بتوزيع أعداد البحوث المترجمة على سنوات النشر لكل منها:

السنة	عدد الأبحاث
١٩٩٦	١
١٩٩٧	٠
١٩٩٨	٠
١٩٩٩	٠
٢٠٠٠	٠
٢٠٠١	٠
٢٠٠٢	١
٢٠٠٣	٢
٢٠٠٤	٢
٢٠٠٥	١
٢٠٠٦	٢
٢٠٠٧	٧
٢٠٠٨	٩

٢٠٠٩	١٠
٢٠١٠	٤
٢٠١١	٦
٢٠١٢	١٨
٢٠١٣	١٢
٢٠١٤	١٩
٢٠١٥	٢٠
٢٠١٦	٢٣
٢٠١٧	٥
٢٠١٨	١



وفيما يأتي جدول بالمواضيع البحثية الرئيسية والفرعية التي صنفتها اللجنة للبحوث المختارة:

الموضوع الرئيسي	الموضوع الفرعي	ترجمة أبحاث المسح	الترجمة الطويلة نوع أ	الترجمة القصيرة نوع ب
اللغة	الصرف		٥	٨
	النحو		٣	٦
الموارد	المعاجم الآلية		٣	٤
	المكانز والذخائر اللغوية	١	١	٦
	المدونات		٣	٢
	الأنطولوجيا		٢	٨
	شبكات الكلمات		٢	٤
التطبيقات	التعرف الصوتي وتوليد الكلام		٤	١١
	القارئ الآلي	١	١	٣
	التشكيل الآلي		٥	٣
	الترجمة الآلية	١	٢	٣
	التحليل الدلالي		١	٦
	البحث في النصوص		٠	٦
	السؤال والجواب		٤	٢
	تلخيص النصوص	١	١	٦
	تحليل الرأي	١	٤	٢
	التعرف على أسماء الأشياء		٢	١٨
	التعليم والتعلم الآلي		٢	٢

٤-٢-٣ الخلاصة

لا تفتقر حوسبة اللغة العربية إلى الباحثين الجادين، لكنها تفتقر إلى عناصر أساسية حتى تتحقق حوسبتها، وحتى تصل بعد ذلك إلى مستوى حوسبة اللغة الإنجليزية. وأهم هذه العناصر:

الإطار النظري اللغوي الحديث الذي يصف اللغة العربية.

الموارد اللغوية المحوسبة الأساسية لحوسبة اللغة العربية.

البيئة العملية والمادية اللازمة لتحويل نتائج البحوث والدراسات إلى منتجات.

وغياب هذه العناصر يعني أن مشروع حوسبة اللغة العربية لم يزل في بداياته، رغم كم الأبحاث، وتعدد المشاريع والمحاولات. ولذلك فالمنتجات التي تقوم على حوسبة اللغة العربية قليلة جدا، وهي تقتصر على عدد محدود من المجالات التيى لا تغطي جوانب التقدم التقني لتطبيقات حوسبة اللغة، أما الجهات التي تستطيع إنتاج هذه التطبيقات فلم تزل محصورة في الشركات الكبيرة.

ولا شك أن التحديات التي تواجه حوسبة اللغة العربية لكبيرة، لا تستطيع جهة واحدة مواجهتها منفردة دون مشاركة حقيقية من جهات متعددة.

٤-٣ التوصيات

إن استمرار غياب التطبيقات التي تقوم على حوسبة اللغة العربية يعني أن استخدامنا للتقنية سيكون باللغة الإنجليزية، وهذا الاستخدام سيبدأ من الصغر في مراحل التعليم المختلفة، وسيمتد مع تقدم العمر إلى كل نشاطات الحياة. وإذا استمر هذا الغياب لأكثر من جيل فلا بد أن اللغة الأم ستنحصر ويتراجع استعمالها، إلى أن تصبح في أفضل الأحوال لغة للعبادات لا علاقة لها بالواقع.

وإذا كان تطور حوسبة اللغة الإنجليزية يرتبط بخصوصيات تتعلق بهذه اللغة، منها

انتشارها العالمي، وغنى الدول والجهات التي تتبنى مشاريع حوسبتها، فإن للغة العربية خصوصيات لا بد من أخذها بعين الاعتبار عند النظر في موضوع حوسبتها. هذه الخصوصيات منها ما يتعلق بذات اللغة العربية، ومنها ما يتعلق بالبيئة التي تحيط بها، أي بالواقع الثقافي والاجتماعي والاقتصادي والسياسي لأهلها. لذلك فالتوصيات التي نخرج بها من مراجعتنا لواقع حوسبة اللغة العربية تراعي هذا الواقع وهذه الخصوصيات.

وفيما يأتي توصيات اللجنة في ضوء العناصر التي وردت في الخلاصة السابقة:

وضع خارطة طريق للوصول إلى إطار نظري جديد لوصف اللغة العربية، يستفيد مما وصل إليه البحث اللساني الغربي، ويراعي متطلبات حوسبة اللغات الطبيعية، وخصوصيات بنى اللغة العربية. ولخطر هذا العمل فإن مجرد وضع تصور له يعد اختراقا علميا يتطلب الكثير من البحث والجهد.

وضع خارطة طريق لحوسبة اللغة العربية، فإن تشتت الجهود لحوسبة اللغة العربية يقتضي وضع إطار لهذا العمل، لتصنيف الأعمال المختلفة، وقياس مدى الإنجاز، وتجميع الجهود، وتراكم المعرفة، وتجنب التكرار.

تصنيف مشاريع البحث والتطوير التي جرت حتى اليوم في حوسبة اللغة العربية إلى موارد أو تقنيات أساسية، وتقنيات لها حقوق الملكية الفكرية. ودعوة الباحثين والشركات إلى إتاحة الموارد والتقنيات الأساسية حتى يمكن استخدامها دون قيود في عمليات التطوير المستقبلية.

إيجاد مؤسسات غير ربحية تجمع الموارد الأساسية لحوسبة اللغة العربية وتتيحها لمن يشاء.

وضع قائمة بمجموعة من مشاريع الموارد الأساسية، والدعوة إلى تمويل هذه المشاريع وإيجاد فرق عمل تقوم على تطويرها، ثم إتاحتها لمن يشاء. ومن أهم هذه المشاريع:

مشروع شبكة الكلمات العربية، وما يتفرع عنها من انطولوجيات.

وسم المدونات المشكولة.

المحللات اللغوية: الصرفية والنحوية والدلالية والمقامية.

الدعوة إلى إيجاد مؤسسات مستقلة تقوم على قضية حوسبة اللغة العربية. فالمرحلة التي وصل إليها العلم والتقنية لا تحتمل ترك مسألة حوسبة اللغة العربية للتجريب الأكاديمي والتجاري. ولا بد من وجود مؤسسات قائمة على موضوع حوسبة اللغة العربية. على أن تحظى هذه المؤسسات بالدعم الرسمي والمالي لنشاطاتها، لتقوم بوضع الأطر، وتنسيق النشاطات، وتوفير الدعم المادي لعمليات حوسبة اللغة العربية، وما يرافقها من بحث علمي. ثم إتاحة ما وصلت إليه مشاريع البحث والتطوير لمن يريد أن يبني عليه تطبيقات تقنية.

الملحق 1
المصطلحات

لمصطلح الإنجليزي	الترجمة المقترحة	وصف المصطلح
Abstractive Summarization	تلخيص تجريدي	تقنية التلخيص التجريدي تشبه إلى حد ما عملية "إعادة الصياغة" للنص . فالنصوص الموجزة باستخدام هذه التقنية تكون ملخصات أكثر تكثيفا. لكنها أكثر صعوبة من تقنيات التلخيص الاستخراجي.
Accuracy	الدقة	هي أحد أشهر مقاييس الحكم على الأنظمة، وتعبر عن دقة النظام في تصنيف البيانات مثلا، وقد يستخدم في تحديد دقة أنظمة التعرف.
Accusative	مفعول به	الاسم المنصوب الذي يدل على من وقع عليه فعل الفاعل، وقد يكون اسمًا مفردًا أو جملة أو شبه جملة أو غير ذلك.
Acoustic Model	النموذج الصوتي	هو نموذج يستخدم في التعرف التلقائي إلى الكلام لتمثيل العلاقة بين الإشارة الصوتية والصوتيات أو الوحدات اللغوية الأخرى التي تشكل الكلام.
Active Participle	اسم الفاعل	اسم مشتق يدل على من قام بالفعل، على سبيل المثال اسم الفاعل من كتب "كاتب ".
Adjacency Constraints	ضوابط التجاور	ضوابط التجاوز أو عدم التوافق(**incompatibility**) مثل قاعدة أنه لا يمكن أن يتتابع حرفا جر.
Affirmative	التأكيد	كلمة مرتبطة بجملة أو فكرة تأكيدية دون أي شك.
Affixes	زوائد	هي الحروف الزائدة التي إما أن تأتي في أول الكلمة وتسمى السوابق (**prefixes**) أو تأتي في آخر الكلمة وتسمى اللواحق (**suffixes**) أو التي تأتي في داخل الكلمة وتركيباتها وتسمى.(**infixes**)
Agglutination	تلصيق	هي تلك العملية في اللغة التي تسمح بإضافة سلاسل للكلمة دون تغيير الصوت أو المعنى ومن الأمثلة على اللغات التي تدعم هذه الخاصية: اللغة العربية والتركية.

Algorithm	**خوارزمية**	**مجموعة من الخطوات المرتبة والمتسلسلة منطقيا لحل مشكلة ما، ولا تشترط لغة معينة لكتابتها، وقد سميت بهذا الاسم نسبة إلى العالم الخوارزمي.**
Aligned Arabic/ English texts	**النصوص العربية/ الإنجليزية المُتحاذية**	**يقصد بها محاذاة النصوص في الوثائق بحيث تضم الوثيقة النصوص العربية وما يقابلها من النصوص الإنجليزية.**
Alphabet	**الحُروف الهجائية**	**هي مجموعة أصوات ذات نظام خاص في لغة ما تعرف بأنها حروف البناء التي تتكون منها الكلمات، مثلا عدد حروف اللغة العربية ٢٨ حرفا والإنجليزية ٢٦ حرفا**
Alteration rule	**قانون التصريف**	**قاعدة التناوب تزود اللغويين بالبيانات التي تسمح لهم بتحديد الأصوات والأشكال المكونة لصوتيات اللغة والمورفولوجيات اللغوية.**
Ambiguity	**غموض**	**الغموض هو نوع من عدم التيقن من المعنى الذي تحمل فيه الكلمة عدة تفسيرات معقولة. وبالتالي فإن الغموض من سمات أي فكرة أو بيان المقصود الذي لا يمكن حله بشكل نهائي وفقا لقاعدة محددة.**
Analogous	**مُتناظرة**	**متناظرة مصطلح يطلق على عنصرين يحتويان عناصر متشابهة، على سبيل المثال الحاسوب يشبه الدماغ.**
analyzable	**قابلة للتحليل**	**صفة تطلق على بيانات أو معلومات قابلة للتحليل لبناء نتائج أو افتراضات.**
Annotated Corpora	**الذخيرة اللغوية الموسومة**	**هي تلك المجموعات من البيانات التي يتم وسم كلماتها بمعلومات لغوية إضافية تسمي "الوسوم". ويمكن أن تكون هذه المعلومات مثل النوع الصرفي أو النحوي للكلمة.**
Annotation	**وسم**	**وسوم التعريف (مثلا ، الوظيفة ، الشرح ، العلامات، النوع الصرفي) المرفقة بالكلمة**

Annotation of Corpora	وسم الذخيرة اللغوية	هي عملية وسم النصوص بعلامات تدل على تصنيف كلماتها صرفيا ونحويا
Applied Linguistics	اللغويات التطبيقية	هو فرعٌ من فروع اللسانيات "أي علم اللغة العام"، وهذا الفرعُ يعنى بتطبيق النظريات اللغوية ومعالجة المشكلات المتعلقة باكتساب اللغة الأولى والثانية وتعليمها. كما يعنى هذا الحقل بالتحليل التقابلي بين اللغات للاستفادة منه في تحسين ظروف تعلم اللغات وتدريسها
Approach	منهجية أو أسلوب أو مقاربة	هي الطريقة أو النهج المتبع لحل مشكلة ما.
Approximate String-Matching	مطابقة تقريبية بين النصوص	هو أسلوب البحث في النصوص عن سلاسل الحروف التي تطابق نمطا ما بشكل تقريبي (مقابل إيجاد سلسلة متطابقة تماما).
Arabic Character	الحرف العربي	هي مجموعة الحروف التي تشكل النص العربي كما هو . وهي مكتوبة من اليمين إلى اليسار وتشمل ٢٨ حرفا.
Arabic Corpora	الذخائر اللغوية العربية	مجموعة من النصوص العربية تكون قد جمعت لأغراض البحث والتحليل الآلي للنصوص.
Arabic Electronic Text	النصوص الإلكترونية العربية	هي النصوص العربية المخزنة على شكل ملفات في جهاز الحاسوب أو على شبكة الإنترنت بخلاف النصوص المطبوعة على الورق.
Arabic Lexicon Ontology	أنطولوجيا المُعجم العربي	هو علم توصيف المعجم العربي بتجميع الكلمات العربية في مجموعات ذات علاقات بينها مثل الترادف والتضاد وعلاقات الارتباط والتصريف والاشتقاق.
Arabic Named Entity (ANE)	الأسماء العربية للأشياء	هي الأسماء التي تطلق على الأعلام في اللغة العربية مثل أسماء الأشخاص، أسماء الأماكن والمنظمات وغيرها.

Arabic Named Entity Recognition (ANER)	التعرف إلى الأسماء العربية للأشياء	هي برامج التعرف الآلي التي بإمكانها استخلاص الأسماء (أسماء الأعلام والأماكن والمنظمات) من مجموعات النصوص العربية.
Arabic Natural Language Processing (ANLP)	معالجة اللغة العربية حاسوبيا	مجموعة من البرمجيات والأدوات لتحليل وتفسير اللغة العربية تفسيرًا آليًا.
Arabic Ontology Model	نموذج أنطولوجي عربي	نموذج وصفي آلي لتمثيل المعجم العربي.
Arabic Opinion Mining	التنقيب عن الرأي العربي	أحد فروع علم الحاسوب ويهتم بتحليل مشاعر الناس وآرائهم ومواقفهم وتقييماتهم المعرب عنها باللغة العربية.
Arabic Parsers	المُحللات اللغوية العربية	هي البرامج التي تقسم الكلام حسب التركيب مثل الكلمة وشبه الجملة والجملة وتحلل هذه التراكيب صرفيا ونحويا ودلاليا.
Arabic Script	النص (الخط) العربي	هو نظام الكتابة المستخدم لكتابة اللغة العربية وهو مكتوب من اليمين إلى اليسار.
Arabic Stemmers	مُحللات الجذوع العربية	هي عملية تحديد الكلمة العربية باستبعاد الزوائد منها ، استنادا إلى جذر الكلمة. وقد طور المبرمجون برامج آلية للقيام بهذه المهمة.
Arabic Text	النَص العربي	مجموعة من الكلمات مكتوبة باللغة العربية.
Arabic Text-to-Speech (ATTS)	تحويل النص العربي إلى كلام منطوق	عملية تحويل النصوص المكتوبة باللغة العربية إلى كلام منطوق.
Arabic Topic Detection	اكتشاف موضوع النص العربي	هي القدرة الآلية على تحديد موضوع النص المكتوب اعتمادا على كلمات مفتاحية داخل النص.
Arabic Traditional Grammar (ATG)	القواعد النحوية العربية التقليدية	هي قواعد اللغة العربية التراثية المعروفة.

ويكيبيديا العربية هي مشروع موسوعة باللغة العربية، مبنية على الويب، ذات محتوى حر، تشغلها مؤسسة ويكيبيديا، التي هي منظمة غير ربحية. ويكيبيديا هي موسوعة يمكن لأي مستخدم تعديل وتحرير وإنشاء مقالات جديدة فيها.	الويكيبيديا العربية	Arabic Wikipedia
هو برنامج أو تطبيق يستخدم لاستخلاص الكلمات العربية من نصوص مكتوبة بخط اليد أو من صور تتضمن كلمات، ويستخدم في برامج التعرف الآلي على الحروف.	راصد الكلمة العربية	Arabic Word Spotter
قاعدة بيانات معجمية تجمع كلمات اللغة العربية وصفاتها والعلاقات الصرفية والنحوية والدلالية بينهما.	شبكة الكلمات العربية	Arabic WordNet (AWN)
استخدام الحروف العربية لتشكيل الكلمات العربية التي تتجمع لتكون النصوص العربية.	الكتابة العربية	Arabic Writing
"التركيب: تأليف الشيء من مكوناته البسيطة، ويعني اصطلاحا ضمُّ أو رَصْفُ اسمٍ إلى جانب اسمٍ، أو فعلٍ إلى جانب اسمٍ؛ ليُكوِّنا كلامًا مفيدًا يؤدي وظيفته الاتصالية ويَقبله المتلقي.	البنية أو التركيب	Architecture
تقنيات حسابية لتجميع الكلام استنادا إلى نماذج من المسالك الصوتية والبشرية وعمليات التعبير التي تحدث أثناء عملية التكلم.	التركيب اللفظي	Articulatory Synthesis
هو خصائص معينة تتسم بها البرامج الحاسوبية تجعلها تحاكي القدرات الذهنية البشرية وأنماط عملها. من أهم هذه الخاصيات القدرة على التعلم والاستنتاج ورد الفعل على أوضاع لم تبرمج في الآلة.	الذكاء الاصطناعي	Artificial Intelligence

Artificial Neural Network	شبكة عصبية صناعية	نظم الحوسبة المستوحاة في تصميمها من الشبكات العصبية البيولوجية التي تشكل أدمغة الكائنات الحية، وأهم وظائفها حفظ المعلومات والبناء عليها من خلال استخدامها لتحليل بيانات واكتساب معلومات جديدة وأبرز تطبيقاتها أنظمة تعلم الآلة.
Aspect	وحالة الفعل	حالة الفعل كالفعل التام، أو الفعل المستمر، وهي خاصية توضح زمن الفعل
Assimilation	إدغام	عملية صوتية يصبح من خلالها صوت واحد أشبه بصوت آخر قريب. ويمكن أن يحدث هذا إما داخل كلمة أو بين الكلمات، وفي العادة هناك حروف محددة للإدغام.
Attribute	خاصية	الخاصية هي السمة التي تستخدم لوصف الأشياء، على سبيل المثال : اللون هو سمة لخلفية صورة.
Augmented	مزيد	مقطع مضاف إلى بداية الكلمة
augmented Letters	الحروف الزائدة	هي الحروف التي تضاف إلى الكلام دون إضافة معنى جديداً و إنما تؤكد و تقوي المعنى العام في الجملة كلهاً فشأنها شأن كل الحروف المؤكدة تفيد توكيد المعنى العام للجملة
Automata (automaton)	آلي	آلة تقوم بعمل يشبه فعل البشر، مثلاً نمذجة قاعدة لغوية معينة تستعمل للتوليد والتحليل
Automated Dictionary	قاموس قابل للقراءة الآلية	ا لقا مو س ا لذ ي يعمل با لحد الأدنى من التدخل البشري ؛ لاستخراج معاني ولفظ المصطلحات.
Automatic Arabic Speech Segmentation	تقطيع الكلام العربي آليا	عملية تقسيم النص العربي المكتوب آلياً إلى وحدات ذات معنى، مثل الكلمات أو الجمل

هو خطوة في مجال معالجة اللغات الطبيعية تتعلق باستخراج المصطلح آليا كأساس للتطبيقات المختلفة، مثل بناء قاموس خاص ، وأنظمة استرجاع المعلومات ، والترجمة الآلية، والفهرسة، وخلق خريطة المعرفة وتطور المعرفة وانتشارها.	استخراج آلي للمصطلحات العربية	Automatic Arabic Terminology Extraction
هي عملية استغلال تطبيقات الحاسوب والذكاء الاصطناعي لاستخلاص بعض من كل، مثلا الاستخلاص الآلي للأسماء من النصوص العربية.	الاستخراج التلقائي	Automatic Extraction
توظيف الحاسوب وتطبيقات الذكاء الاصطناعي لاستخراج الكلمات التي تم اقتراحها بناء على مواصفات معينة	الاستخلاص الآلي للمفردات المرشحة	Automatic Extraction of CandidateTerms
هي قائمة محوسبة تجمع كلمات في لغة معينة، على نسق منطقي معين، وتهدف إلى ربط كل كلمة منها بمعناها، وإيضاح علاقتها بمدلولها	معجم آلي	Automatic Lexicon
القيام بتحليل النص أو الكلام وتمثيله بشكل قابل للتعامل مع الحاسب ويشبه ذلك ما يقوم به القارئ أو المستمع البشرى.	معالجة اللغات الطبيعية التلقائية	Automatic Natural Language Processing
النهج الذي يستخدم قواعد مكتوبة لوسم النصوص، وهذا النهج يعتمد علي القاموس أو المعجم للحصول على الإشارات الممكنة لكل كلمة لتكون موسومة.	وسم الكلمات التلقائي المستند للقواعد	Automatic Rule-Based Tagging
هو المجال الفرعي من علوم اللغويات الحسابية التي تعنى بتطوير المنهجيات والتكنولوجيات التي تمكن من التعرف وترجمة اللغة المنطوقة في النص من أجهزة الكمبيوتر.	تعرف آلي إلى الكلام	Automatic Speech Recognition (ASR)
هو عملية تلخيص وثيقة نصية بواسطة أحد البرامج، من أجل إنشاء ملخص مع الاحتفاظ بالنقاط الرئيسية في الوثيقة الأصلية.	التلخيص الآلي	Automatic Summarization

Automating	أتمتة	يمكن تعريفها بأنها التكنولوجيا التي يتم بها تنفيذ عملية أو إجراء دون مساعدة بشرية.
Band Pass Filter	مرشح نطاق الذبذبات	هو جهاز يمرر الذبذبات ضمن نطاق معين ويرفض أو يخفف الذبذبات خارج هذا النطاق. والهدف الرئيس منه هو فلترة البيانات قبل معالجتها.
Base Phrase Chunker (BPC)	أداة تقسيم النص إلى عبارات أساسية	مكون البرنامج الذي يأخذ بيانات الإدخال (النص في كثير من الأحيان) ويبني منه بنية البيانات المكونة من جمل وعبارات.
Base Phrase Chunking	تكديس العبارات الأساسية	العملية التي يتم فيها تجميع سلسلة من الكلمات المجاورة معا لتشكيل العبارات النحوية.
Baseline	خط الأساس	هو الخط التخيلي الذي تنتظم فيه الحروف العربية عند كتابتها لتكوين الكلمات.
Bidirectional	ذو اتجاهين	هو أسلوب معالجة النصوص ذو اتجاهين: من اليمين إلى اليسار وبالعكس
Bigram	تركيب مزدوج	زوج من الأشياء، عادة ما يكون زوجاً من الفئات المعجمية، ويستخدم ن-غرام كطريقة لاستخراج الميزات وتقليص عددها.
Bilingual	ثنائي اللغة	هو القدرة على التحدث بلغتين مختلفتين بطلاقة كاملة. أو هو ما يحتوي على نفس المعلومات بلغتين مختلفتين.
Bilingual Corpus	ذخيرة ثنائية اللغة	ذخيرة لغوية تحتوي على نصوص من لغتين: على سبيل المثال مجموعة من النصوص الأصلية في لغة وترجماتها إلى لغة ثانية.
Binarisation	التمثيل الثنائي	هي عملية تحويل النقاط المكونة للصورة إلى إحدى حالتين إما أبيض أو أسود ليمكن تحليلها لأغراض عمليات التعرف إلى الحروف

Blind Test	فحص عشوائي	طريقة تستخدم للكشف عن المشاكل الأمنية التي تحدث نتيجة الأخطاء في الترميز أو الثغرات الأمنية سواءً كانت في النظام أم في البرامج المستخدمة أم الشبكات من خلال إرسال أو إدخال كمية ضخمة جداً من البيانات العشوائية إلى تطبيقات مختلفة أو إلى الإنترنت
Borrowed Word	الكلمات المعرَّبة	هي مجموعة من الكلمات الدخيلة على اللغة التي تم اقتباسها من لغات أخرى دون ترجمتها الحرفية.
Breakdowns	فواصل	علامات التنقيط وتستخدم غالبا في اللغة العربية للفصل بين الجمل القصيرة وللدلالة على أن الكلام مستمر بعدها.
Brevity	إيجاز	هو استخدام مبسط ودقيق للكلمات عند استخدامها في الحديث أو الكتابة.
Broad Coverage Language Resource	مورد لغوي شامل التغطية	مورد لغوي معجمي أو نحوي يحتوي كميات كبيرة من المفردات والقواعد التي تغطي طيفا واسعا من مفردات اللغة ومستوياتها
Broken Plural	جمع التكسير	هو صيغة الجمع غير المنتظمة و يتم تشكيلها عن طريق تغيير نمط الحروف الساكنة والحروف المتحركة داخل شكل المفردة.
Buckwalter Arabic Morphological Analyzer (BAMA)	محلل باكوولتر الصرفي للغة العربية	محلل يستخدم منهجاً مبنياً على حوسبة المعجم، حيث يقوم على بناء القواعد الصرفية والإملائية داخل المعجم نفسه بدلاً من أن تكون محددة من حيث القواعد العامة التي تتفاعل لتحقيق الناتج.
Canonical Verbs	الكلمة المجردة	الكلمة المفتاحية في المعجم
Capitalization	الكتابة بأحرف كبيرة أو ابتداء الكلمات بأحرف كبيرة	هي عملية يتم من خلالها استخدام الأحرف الكبيرة في بداية الكلمة في مواضع معينة.

Case Marker	العلامة الإعرابية	وحدة صرفية تلحق بالكلمات (الأسماء تحديداً) لبيان الحالة الإعرابية للكلمة (مرفوع، منصوب، مجرور)، وتكون حركة، أو حروفًا
Case-Ending	حركات آخر الكلمة	تشكيل آخر الكلمةٍ وهي حالة خاصة مرتبطة باللغة العربية لفهم معنى الكلام.
Categorical	تصنيفي	يستخدم فئة أو فئات، الشكل التصنيفي لمشكلة معينة ويهدف إلى تقسيم عناصرها إلى فئات حسب أوجه الشبه.
Categorical Grammar	نحو تصنيفي	مصطلح يستخدم لعائلة من الشكليات في بناء اللغة الطبيعية بدافع مبدأ التركيب.
Characterizing Measurements	قياسات دالة	القياسات التي تصف طبيعة أو ميزات شئ ما، وتستخدم هذه المقاييس للحكم على نتائج أداء الأنظمة.
Chatbot	روبوت الدردشة	هو برنامج حاسوبي يقوم بإجراء محادثة عن طريق الأساليب السمعية أو النصية.
Chunk	قطعة	يطلق هذا المصطلح على قطعة كبيرة أو قصيرة (من خشب أو خبز أو غيره) أو مقدار وافر من كلّ، كذلك تطلق على مَعْلومَةٌ يُمْكِنُ اخْتِزانُها.
Circumflex	علامة تشكيل	علامات تشكيل تستخدم في اللغات اللاتينية واليونانية
Classical Arabic	اللغة العربية الكلاسيكية	اللغة العربية القديمة التي تداولها العرب في عصر ما قبل الإسلام، والعصرين الإسلاميين الأموي والعباسي
Classification	تصنيف	نهج لبناء نماذج التصنيف من مجموعة بيانات المدخلات. ويعبر عن فصل مجموعة من البيانات إلى فئات اعتمادا على معايير معينة.
Classifier	مصنف	يستخدم لإنشاء توقع لمدخلات معينة إلى فئات اعتمادا على خاصية لتلك المدخلات في سياقها.

Clustering	عنقدة	التجميع هو مهمة تجميع مجموعة من العناصر بطريقة تكون فيها العناصر في نفس المجموعة (تسمي عنقودا) وتكون أكثر تشابها ببعضها عن تلك الموجودة في مجموعات أخرى.
Clusters	عناقيد / مجموعات منفصلة	هي مجموعة من الأشياء المتشابهة أو التي تحدث معا بشكل وثيق وتندرج تحت نفس الفئة.
Coarticulation Effects	أثر دوران الحروف	الوضع الذي يتأثر فيه صوت الكلام المعزول مفاهيمياً ويصبح أشبه بصوت الكلام السابق أو اللاحق
Collaborative Annotation	الوسم التعاوني	هو استراتيجية وسم النصوص باعتماد الأسلوب الآلي ثم اليدوي الجماعي.
Colloquial	اللهجة العامية	هي اللغة الشفوية المستخدمة في محادثة عادية غير رسمية وتكون مالوفة ضمن مجتمع واحد.
Combination Methods	طرق التشكل	هي الطرق التي تنتج من جمع طريقتين أساسيتين لحل مشكلة ما، وتسمى أيضا الطرق الهجينة، التي تهدف إلى رفع كفاءة النظام عن الطريقة التقليدية.
Common Errors Dictionaries	قواميس الأخطاء الشائعة	هو قاموس يقوم على جمع الأخطاء التي يقع بها مستخدمو اللغة وتصحيحها.
Compiler	مترجم	برنامج حاسوبي يحول رمز الحاسوب المكتوب بلغة برمجة واحدة (لغة المصدر) إلى لغة برمجة أخرى (اللغة المستهدفة).
Complementizer Phrase	الجملة الخبرية	هي الجملة التي تحمل خبرا يحتمل الصدق والكذب لذاتها، والتي يمكننا الحكم عليها إذا كانت منافية للواقع.
Compositionality	تركيبية	استخراج دلالة الكلمة من مكوناتها الصرفية والنحوية
Comprehensive	شامل، محيط	هو الإلمام بموضوع ما أو شيء من كل جوانبه أو عناصره

Computational Linguistic Sciences	علم اللسانيات الحاسوبية	علم بيني يعنى بنمذجة اللغات الإنسانية بطرق إحصائية أو قائمة على قواعد محددة وذلك لتهيئة اللغة للتعامل معها من خلال الحاسب. وهو الدراسة العلمية للغة الإنسانية من منظور حسابي حيث يهتم متخصصو هذا العلم ببناء نماذج حسابية لمختلف الظواهر اللغوية.
Computational Morphology	الصرف الحاسوبي	هو أحد العلوم اللغوية الحاسوبية التي تتعامل مع معالجة الكلمات والأشكالها المكتوبة أو المنطوقة
Computer Assisted Language Learning	تعليم اللغة بمساعدة الحاسوب	هو نهج لتدريس اللغة يستخدم فيها الحاسوب كوسيلة للمساعدة في عرض وتقوية وتقييم المواد التي يمكن تعليمها، وعادة ما تتضمن عنصرا تفاعليا كبيرا.
Computer Assisted Language Learning System (CALL)	نظام تعلم اللغة بمساعدة الحاسوب	هو نهج لتعلم اللغة يستخدم فيها الحاسوب كوسيلة للمساعدة في عرض وتقوية وتقييم المواد التي يمكن تعلمها، وعادة ما تتضمن عنصرا تفاعليا كبيرا.
Computerized Language Tools	أدوات اللُغة المُحوسَبة	هي مجموعة من الأدوات التي تساعد في فهم لغة جهاز الحاسوب لتسهيل عملية التواصل معه.
Computerized Speech Laboratory (CSL)	مختبر الكلام المحوسب	هو أحد المرافق التي توفر الظروف الكاملة لاختبار ومعالجة الكلام و الإشارات (البرمجيات والأجهزة) المستخدمة في البحوث.
Concatenation	إلصاق	خاصية صرفية للغة العربية تعمل على إلصاق الزوائد والسوابق واللواحق بالكلمات المجردة لإنتاج أو إشتقاق كلمات جديدة
Concatenative Synthesis	تركيب تسلسلي	هو تقنية لتجميع الأصوات من خلال تسلسل عينات قصيرة من الصوت المسجل (تسمى وحدات)
Concept Extraction	استخلاص المفاهيم	هي عملية يتم من خلالها استخلاص المفاهيم من نصوص محددة.

Concordance	فهرس مواقع الكلمة	تحلّل برامج الفهرسة الآليّة النّصوص والمدوّنات الضّخمة للبحث عن كلمات أو جذوع أو جذور يحتاجها المستخدم وتُقدّم في شكل سياقات دنيا تحدد مواقع الكلمة في النص.
Confusion Matrix	مصفوفة اللبس	مصفوفة حجمها ٢*٢ تهدف للتعبير عن دقة الأنظمة في اتخاذ القرار السليم في التصنيف اعتمادا على عدد العناصر المسترجعة وعدد العناصر المرتبطة.
Conjunctions	حروف العطف	هو أحد حروف المعاني، وجزء من أجزاء الكلام يستخدم لتوصيل العبارات أو الجمل أو تنسيق الكلمات في نفس العبارة.
Consonants	الحُروف الساكنة	صوت الكلام الذي هو مفصل مع إغلاق كامل أو جزئي من القناة الصوتية ويمكن دمجها مع حرف علة لتشكيل مقطع.
Consonants–Half Vowels	أشباه حروف العلة في اللغة الإنجليزية (w & y)	هو الصوت الذي يشبه صوتيا حروف العلة لكنه من الحروف المتحركة وليست الصامتة.
Constituent Structure (C-Structure)	في التحليل النحوي الهرمي أو الشجري (وحدة وظيفية قد تكون كلمة أو جملة)	هو توضيح خصوصية كل واحدة من مفردات الجملة، بملاحظة نوع الكلمة وصيغتها وتمثيلها على شكل شجرة أو هرم.
Context Free Grammar	نحو السياق الحر	نحو صوري على صورة مجموعة من قواعد التوليد النحوي
Contexts	سياقات	هي المعطيات التي يجري فيها الكلام أو الأجزاء السابقة واللاحقة لجملة أو كلمة مكتوبة بحيث يوضح معناها.

Continuation Classes	برمجية المتابعة	يمكن استخدام مصطلح الاستمرارية للإشارة إلى "فئات الاستمرارية"، التي هي مقاطع تعطي لغة البرمجة القدرة على حفظ حالة التنفيذ في أي نقطة والعودة إلى تلك النقطة في وقت لاحق في البرنامج ، وربما عدة مرات.
Contour	محيط الشكل	هو الخط المحيط برسم الحرف لأغراض التعرف إلى الحروف، أو الخط المحيط بشكل موجة المقطع الصوتي لأغراض التعرف إلى الكلام.
Contrastive Inter-Language Analysis	تحليل التباين بين اللغات	طريقة في البحث عن نصوص لتعلم الآلة. تصميمها يساعد في الكشف عن مجموعة واسعة من السمات المميزة للغة المتعلم وتقييم درجة التعميم في جميع مجالات التعلم.
Conversational Agent	أداة حوار أو محادثة	هو نظام كمبيوتر يهدف إلى التحدث مع الإنسان وقد استخدمت أنظمة الحوار هذه النص والكلام والرسومات، والإيماءات وغيرها من وسائل الاتصال على كل من قنوات الإدخال والإخراج.
Convolution	التواء	وظيفة تقوم برسم صف من المتواليات على شكل متوالية من الصفوف
Co-Occurrence	التكرار	هو مصطلح لغوي يعني تكرار حدوث فترتين من النص جنبا إلى جنب مع بعضها في ترتيب معين.
Corpus	ذخيرة لغوية	مجموعة النصوص التي تغطي مجالاً ما أوعدة مجالات
Corpus Acquisition	بناء الذخيرة اللغوية	بنك معطيات تُخزّن فيه المدوّنة في شكل نصوص محوسبة تُدخل بشتّى الأشكال (نص امتداد doc أو pdf أوhtml أو rtf) يديرها محرّك بحث بوساطة بروتوكول يربط بينها جميعا.

Cosine Similarity	تشابه جيب التمام الزاوي	هو مقياس يستخدم لتحديد درجة التشابه بين اثنين من المتجهات غير الصفرية عن طريق تحديد جيب التمام للزاوية التي تقع بينهما.
Critics	نقاد	هم الأشخاص الذين يحكمون على مزايا المصنفات الأدبية أو الفنية بحكم مهنتهم.
Crowdsourcing	حشد الموارد	هي العملية التي يتم من خلالها الحصول على المعلومات أو المدخلات في مهمة أو مشروع ما من خلال الاستعانة بعدد كبير من الناس، سواء المدفوعة أو غير المدفوعة، عن طريق الإنترنت.
Cue or Trigger Words	الكلمات المُحفّزة أو القرينة	هي مجموعة من الكلمات التي تستخدم لربط الجمل وغالبا تستهل الجمل وتشير إلى العلاقات الدلالية في النص، مثل في الوقت نفسه، على أي حال، أو من ناحية أخرى وغيرها الكثير.
Cursive	متصلة	هو نمط من أنماط الكتابة ينطوي على ربط الحروف بطًا متناسقا ومتصلا مع بعضه بعضا مما يجعل عملية الكتابة أسرع.
Data Encryption	تشفير البيانات	هي عملية يتم من خلالها تشفير البيانات عن طرق ترجمتها إلى شكل آخر، غالبا بشكل تعليمات برمجية، بحيث لا يستطيع فهم هذه البيانات إلا مجموعة من الناس ممن يمتلكون مفتاح فك التشفير أو كلمة المرور ليتمكنوا من قراءتها.
Data Mining	التنقيب عن البيانات	هو عملية اكتشاف أنماط في مجموعات البيانات الكبيرة واستخراج معلومات جديدة و تنطوي على طرق التعلم الآلي، والإحصاءات، ونظم قواعد البيانات.

Database	قاعدة بيانات	مجموعة منظمة من البيانات الموجودة في الحاسوب، لا سيما تلك التي يمكن الوصول إليها بطرق مختلفة.
Data-Driven Approach	منهجية مشتقة من البيانات	منهجية تعتمد على البيانات لاشتقاق نموذج للتحليل أو المعالجة
Datasets	قوائم البينات	هو عبارة عن مجموعة من البيانات، عادة ما تتوافق المجموعة الواحدة مع محتويات جدول قاعدة بيانات واحد، أو مصفوفة بيانات إحصائية واحدة، حيث يمثل كل عمود من الجدول متغيرا معينا، ويتطابق كل صف مع عضو معين في مجموعة البيانات المعنية.
Decision Tree	شجرة القرار	مخطط لتفرعات يحتوي على القرارات والعلاقات فيما بينها، حيث تمثل بأداة لدعم القرار تستخدم رسما بيانيا شبيها بالشجرة أو نموذجا للقرارات ونتائجها المحتملة ، بما في ذلك نتائج حدث الصدفة ، وتكاليف الموارد ، والأداة المساعدة.
Decision Tree Classifier	مصنف شجرة القرار	مصنف يقوم بتقسيم مجموعة البيانات إلى أقسام أصغر بشكل هرمي، بينما في الوقت نفسه، يتم تطوير شجرة القرارات المرتبطة بشكل متزايد
Deep Belief Networks	الطبقات الداخلية المكونة للشبكة العصبية	هو نموذج رسومي يتكون من طبقات متعددة من المتغيرات الكامنة ("وحدات خفية")، مع وصلات بين الطبقات ولكن ليس بين وحدات داخل كل طبقة.
Definite	معرف	الاسم المعرفة
Definite Article “al”	أل التعريف	هي أداة تستخدم قبل الاسم لتعريفه ولتحديد النحوية له.
Definite Clause Grammar (DCG)	نحو الجمل المقيدة	هو وسيلة للتعبير عن قواعد اللغة، سواء اللغات الطبيعية أو الرسمية، في لغة البرمجة المنطقية. وتسمى بهذا الاسم لأنها تمثل قواعد النحو كمجموعة من العبارات محددة في المنطق من الدرجة الأولى.

Definite Letter	أداة التعريف	هي أداة تستخدم مع الاسم ككلمة مستقلة أو بادئة أو لاحقة لتحديد النحوية له، مثل (أل) التعريف في العربية.
Definiteness	تعريف (مقابل تنكير)	سمة دلالية للعبارات الاسمية تميز بين العبارات التي يمكن تعريفها في سياق معين (عبارات اسم محددة) والكيانات التي لا يمكن تعريفها(عبارات اسم غير محددة)
Dependency Grammar	النحو المعتمد	نظرية الإسناد في النحو كالمسند والمسند إليه، فئة من النظريات النحوية التي تقوم على أساس التبعية للعلاقات، والتبعية هي الفكرة القائلة بان الوحدات اللغوية ، مثل الكلمات ، ترتبط بعضها ببعض بوصلات موجهة.
Derivational	اشتقاقي	الاشتقاق هو تشكيل كلمة جديدة من كلمة أو جذع أو ساق آخر. وعادة ما يحدث ذلك بإضافة أو حذف مقطع . الكلمة المشتقة غالبا ما تكون من فئة كلمة مختلفة عن الأصل.
Derived	مشتق	كلمة جديدة تم اشتقاقها من كلمات أخرى بحيث يتم تعديل المعنى الأساسي للكلمة على سبيل المثال (مدرسة) مشتقة من (درس).
Determiners	المحددات	المحددات هي كلمات تأتي في بداية العبارة وقبل الاسمٍ ومهمتها هي تحديد الأسماء العامةٍ مثل كل ٍ بعض وغيرها كثير.
Detouring Algorithm	خوارزمية التحويل	هي خوارزمية تعتمد على أرقام التحويل لترقيم الخلايا ، والخلايا التي تحمل أصغر الأرقام تتم توسعتها قبل الخلايا التي تحمل أرقامًا كبيرة

Deutsch's Thinning Algorithm	خوارزمية دويج للتنحيف	عبارة عن ثلاث خوارزميات ترقيق للصور: كل واحدة منها للاستخدام مع المصفوفات المستطيلة والسداسية والثلاثية. وسميت بهذا الاسم نسبة إلى صاحب فكرتها "ديوتش".
Diacritical Marks	علامات التشكيل	رموز نصية ثانوية تضاف إلى رمز نصي أساسي (حرف) لتحقيق الفهم والغاية من الكلام. تشمل تلك العلامات التحريك والإعجام والنقط.
Diacritic-Optional Writing System	نظام كتابة بتشكيل اختياري	نظام كتابة يكون فيه تشكيل الكلمات اختياريا حسب رغبة المستخدم.
Diacritics	علامات تشكيل	رموز نصية ثانوية تضاف إلى رمز نصي أساسي (حرف) لتحقيق واحد من أغراض متنوعة. تشمل تلك العلامات التحريك والإعجام والنقط
Diacritization	تشكيل	إضافة رموز ثانوية إلى رمز أساسي (حرف) لتحقيق واحد من أغراض متنوعة
Diacritization Scheme	مستوى التشكيل	يُعدُّ التشكيل أحد الخصائص المهمة للكلمة العربية حيث يفيد التشكيل في تحديد خصائص لغوية أخرى للكلمة، فوجود التشكيل في آخر الكلمة (الفتحة أو الضمة أو الكسرة أو السكون) يفيد في تحديد الحالة الإعرابية للكلمة ووظيفتها في الجملة.
Diacritized Language	لغة مشكولة	هي تلك اللغات التي يستخدم التشكيل في بنائها وأشهر أمثلتها اللغة العربية.
Diacritizer	مُشكّل	ضبط حروف الكلمة بالحركات (إشارات التشكيل: الفتحة والضمة والكسرة والسكون).
Dialectal Arabic (DA)	اللهجة العربية، اللغة الدارجة	اللغة الدارجة، وهي طريقة في استخدام اللغة في مجتمع ما تنزاح عن بعض سمات اللغة الفصيحة وقواعدها.

Dialectal Corpora	الذخيرة اللغوية للهجات	مورد يشتمل على مفردات لهجة في لغة ما تكون مرتبة ترتيبا هجائيا ويمكن أن يكون القاموس ثنائي اللغة مثلا إنجليزي عربي، حيث يعطي مقابل معاني الكلمات الإنجليزية باللغة العربية
Dialectal Usage	استخدام اللهجات	تشير إلى استخدام مجموعة متنوعة من لهجات لغة ما تبعا لمنطقة معينة، وتلك اللهجات تتبع للغة رسمية واحدة.
Dictionary	قاموس أو مُعجم	هو أداة لجمع كلمات لغة ما وتعريفها وشرحها اما بنفس اللغة أو بانشاء قائمة بكلمات لغة ما وما يقابلها في لغة أخرى.
Digital Embedded Systems	أنظمة رقمية مطمورة	أنظمة حماية رقمية تهدف لإخفاء المعلومات كالصور والنصوص والأصوات وتكون مضمنة في وسائط متعددة لغايات الحفاظ عليها.
Digital Ink	الكتابة الرقمية	يشير إلى التكنولوجيا التي تمثل خط اليد في شكله الطبيعي رقمياً حيث يتم وضع جهاز التحويل الرقمي تحت أو فوق الشاشة لإنشاء حقل كهرومغناطيسي و الذي يمكن التقاط حركة القلم الخاص وتسجيل الحركة على الشاشة . ويمكن بعد ذلك حفظ الكتابة اليدوية المسجلة بخط اليد أو ت
Digital Signal Processing	معالجة الإشارات الرقمية	هو استخدام المعالجة الرقمية، مثل أجهزة الكمبيوتر أو معالجات الإشارات الرقمية الأكثر تخصصا، لأداء مجموعة واسعة من عمليات معالجة الإشارات.
Diglossia	ازدواجية اللسان	هو استخدام اكثر من لغة في مجتمع معين. فبالإضافة إلى التنوع اللغوي اليومي أو العام المحلي في المجتمع ، يتم استخدام لغة ثانية مقننة للغاية في حالات معينة مثل الأدب، والتعليم الرسمي، أو إعدادات محددة أخرى، ولكن لا تستخدم في المحادثة العادية.

Dimensionality Reduction	تقليل متجهات العينة	هي عملية تقليل عدد المتغيرات العشوائية من خلال الحصول على مجموعة من المتغيرات الرئيسية.
Disambiguating	فك اللبس	وهو العملية التي يتم من خلالها إزالة الغموض وتوضيح المعنى من جملة غامضة، عبارة، أو وحدة لغوية أخرى.
Discrepancies	اختلافات	عدم التوافق أو التشابه بين حقيقتين أو أكثر
Discrimination	تمييز	التمييز هو اسم يبين المراد من مفرد أو تركيب مبهم قبله، ويأتي -في أغلب حالاته- منكَّرًا منصوبًا، مثل: قرأت عشرين كتابًا .
Distance Time Warping (DTW)	خوارزمية انعطاف زمن المسافة	هي واحدة من خوارزميات لقياس التشابه بين تسلسلين زمنيين، وقد تختلف في السرعة.
Distant Learning	التعلم عن بعد	وهي طريقة للدراسة التي يتم فيها بث المحاضرات أو يتم إجراء الدروس عن طريق المراسلة أو وسائط تقنية أخرى، دون أن يكون الطالب بحاجة إلى حضور مدرسة أو كلية.
Distinctive	مميز	سمة لشخص واحد أو شيء تفيد تمييزه عن غيره.
Document Annotation	وسم الوثائق	الشروح هي التعليقات أو الملاحظات أو التفسيرات أو الأنواع الأخرى من الملاحظات الخارجية التي يمكن إرفاقها بمستند أو بجزء محدد من المستند. من الممكن التعليق علي أي مستند بشكل مستقل ، دون الحاجة إلى تحرير المستند نفسه.
Document Classification	تصنيف الوثائق	هي مشكلة في علم المكتبات والمعلومات وعلوم الكمبيوتر، وتتمثل مهمة التصنيف في تعيين مستند إلى فئة أو مجموعة من الفئات بشكل يدوي أو فكري أو على شكل خوارزمية.

Document Indexing	فهرسة الوثائق	هو وصف أو تصنيف وثيقة حسب مصطلحات الفهارس أو رموز أخرى من أجل توضيح ما هي الوثيقة أو تلخيص محتواها أو زيادة قابليتها للبحث.
Document Web	شبكة الوثائق	هي وثيقة مناسبة لشبكة الويب العالمية ومتصفحات الويب.
Domain	مجال – نطاق	هي نطاقات معينة تشترك بالخصائص نفسها.
Domain Independent System	نظام غير مرتبط بنطاق	هو نظام يوفر مجموعة من الأدوات القادرة على بناء أنظمة تصميم الصور من مكتبة عناصر البرمجيات القابلة لإعادة الاستخدام.
Doubled Letter	حرف مضعف	حرف مشدد وينتج في الأصل من دمج حرفين متماثلين متتاليين.
Dual	المُثنّى	اسم معرب يتم ذكره بدلًا من ذكر اسمين متفقين في اللفظ والمعنى وذلك بزيادة ألف ونون أو ياء ونون، ويرفع بالألف ويُنصب ويجر بالياء
Elliptic Personal Pronoun	الضمير المستتر	هو الضمير المخفيّ فلا يوجد في الجملة صراحة لكن يمكن تقديره و فهمه من سياق الجملة.
Empirical Data	البيانات التجريبية	هي البيانات المستمدة من القياس أو الملاحظة الموثوق بها. كما يمكن تعريفها بأنها وسيلة لكسب المعرفة من خلال المراقبة المباشرة وغير المباشرة أو الخبرة.
Empirical Results	نتائج تجريبية	نتائج تم الحصول عليها بالتجربة. وقد تنتنج من خلال الاختبار العملي لنظام ما.
Enclitic	زائدة نهائية	الزوائد المتصلة في آخر الكلمة كالضمائر
Encode	تشفير	هو تمثيل منهجي للرموز في النص لغرض التخزين والاخفاء وعدم قابلية الاسترجاع لغير المصرح لهم.
English Gloss-Based Approach	المنهجية القائمة على المصطلحات الإنجليزية	هي الطريقة التي تعتمد على المعاني لقياس العلاقات الدلالية بين مفردات اللغة الإنجليزية

English SentiWordnet (ESWN)	شبكة اللغة الإنجليزية لتحليل المشاعر	هي إحدى منهجيات المشاعر التي تعتمد على أساليب استخراج الرأي باللغة الإنجليزية بنجاح.
Entity	كينونة	هو شيء موجود في حد ذاته بشكل ملموس أو مجرد في الواقع، كموضوع أو كائن.
Error Analysis	تحليل الأخطاء	هو علم يدرس أنواع وأسباب الأخطاء اللغوية التي تصنف إلى عدة أنواع ومنها الشكلاً والنوعاً والسبب وغيرها.
Error Tagging Manual	دليل وسم الأخطاء	الدليل الذي يتضمن أدوات يمكنها تمييز الأخطاء واستخراج التعابير الخاطئة في مجموعة من النصوص
Errors	أخطاء	هو إجراء غير دقيق أو غير صحيح. في الإحصاءات "خطأ" يشير إلى الفرق بين القيمة التي تم حسابها والقيمة الصحيحة.
Euclidean Distance	المسافة الإقليدية	المسافة الإقليدية أو المقياس الإقليدي هو الخط العادي المستقيم بين نقطتين في فضاء إقليدس. ويصبح فضاء إقليدس في هذه المسافة فضاء قابلاً للقياس.
Euro WordNet	شبكة الكلمات الأوروبية	هو نظام من الشبكات الدلالية للغات الأوروبية ، على أساس ووردنت، حيث إن كل لغة تطور الشبكة الخاصة بها ولكنها مترابطة مع الروابط بين اللغات المخزنة في المؤشر اللغوي البيني
Evaluation Methods	طرق التقييم	هي المعادلات والصيغ التي يمكن تطبيقها على النظام المقترح لفحص كفاءته من حيث الأداء والسرعة.
Excitation	إثارة	هو نهج دلالي للتحليل. لمحاكاة الطريقة التي يفكر بها الإنسان، واستخدام كل كلمة ابتداء من إثارة المعاني، واتخاذ القرار على أساس ما تعنيه الجملة، بدلا من النظام النحوي لها.

Experimental Method	المنهج التجريبي	وسيلة منهجية للحصول على البيانات للوصول إلى المعرفة بواسطة الرصد أو الملاحظة العلمية بشكل مباشر أو غير مباشر. و يمكن للقانون التجريبي أن يحلل إما بشكل كمّي أو نوعي.
Exploit	استغلال	عملية الاستفادة بشكل أمثل من الموارد المتاحة لنظام أو مشروع ما.
External Features	خصائص خارجية	هي السمات أو الاجزاء البارزة والواضحة للعيان.
Extraction	استخراج	عملية استخراج ميزات خاصة للتعبير عن بيئة بيانية متنوعة.
Extractive Summarization	التلخيص الاستخراجي	التلخيص الاستخراجي يعتمد بشكل أساسي على توليد ملخصٍ لنص ما بدقة عالية، ويمكن تشبيهه باستخراج الأفكار الرئيسة لنص ما.
Extrinsic Evaluation	تقييم في البيئة الحقيقية	طريقة لتقييم أداء مكونات نظم المعالجة الآلية النصية من منظور آثارها على أداء النظام بأكمله.
Facilitate	تسهيل	إعادة تمثيل لمشكلة معينة بهدف تسهيل التعامل معها.
Fast Fourier Transform	تحويل فوريير السريع	واحدة من أشهر الخوارزميات التي تتعامل مع الإشارات وفقا للزمن أو التردد وينسب اسمها لعالم الرياضيات الفرنسي فوريير.
Feature	خاصية	الخصائص التي يتم استخراجها من البيانات الأصلية لتسهيل فهمها واستخدامها.
Feature Extraction	استخلاص الخصائص	عملية استخراج الميزات الهامة التي من شأنها أن تعبر عن البيانات بطريقة أفضل وبوقت أقل.
Feature Modeling Component	وحدة نمذجة الخصائص	الوحدة التي تهتم بنمذجة أساليب خاصة للحصول على الميزات الهامة والفريدة في بيانات محددة.
Feature Point Detection	كشف عن نقطة الخصائص	العملية التي تهتم باكتشاف النقاط ذات الخصوصية والفريدة والتي تُعدُّ سمة أساسية للتعبير عن العينة التي وجدت بها.

Feminine	المؤنث	وهو عكس التذكير وهو كل ما دل على التأنيث مع وجود خصائص يعتقد أنها تخص المرأة، أو هي مجموعة من السمات السلوكية والأدوار المرتبطة عموما بالفتيات والنساء.
Feminine Approach	أسلوب أنثوي	الأسلوب ذو الصفات أو المظهر المرتبط تقليديا بالمرأة
Finite State	حالة منتهية	نموذج رياضي للحساب يتميز بنهاية معروفة.
Finite-State Lexical Transducer	محول معجمي للحالات المنتهية	آلة تعتمد الحالات المنتهية لتوليد الكلمات أو الوحدات المعجمية
First order High Pass Filter	مرشح تمرير الترددات العالية ذو رتبة أولى	مرشح إلكتروني يقوم بتمرير إشارات مع تردد أعلى من تردد قطع معين وتخفيف الإشارات مع ترددات أقل من تردد القطع.
Floating Dictionary	قاموس عائم	قاموس محوسب يعمل على سطح التطبيقات الأخرى
F-Measure	اختبار دقة الفحص	التحليل الإحصائي للتصنيف الثنائي، والنتيجةF1 وهو مقياس لدقة الاختبار. يستخدم للحكم على فعالية النظام.
Foreign Proper Nouns	أسماء أجنبية صريحة	اسم علم أجنبي يشير إلى كيان فريد من نوعه سواء كان اسم شخص أم مؤسسة أم مكان.
Formalization	صياغة	نظام رياضي يعتمد على إعادة صياغة مشكلة ما باستخدام معادلات حسابية لتسهيل الحصول على نتائج أفضل.
Format Synthesis	تركيب الشكل	عدد من الأشكال التي يمكن أن تظهر الخصائص التركيبية للغة، ويوجد شكلان أساسيان؛ الأول يعتمد على ضم التصريفات المختلفة (الأسماء، والأفعال الخ) لخلق كلمات جديدة، والشكل الآخر هو التوليف العلائقي، الذي يعتمد على ضم الجذور منضمة إلى المورفولوجيات

Frames	إطارات	بيئة هيكلية يمكن فيها استخدام فئة من الرموز والأحرف والأرقام بشكل صحيح
Framing	تأطير	يرتبط مفهوم التاطير بالعمليات التنظيمية التي تجرى على مجموعة عملية من البيانات بهدف الإشارة إلى أحداث دلالية معينة مثل تأطير حزم البيانات المنقولة عبر الإنترنت، حيث يكون التأطير عن وضع عنوان المرسل والمستقبل وزمن الإرسال.
Frontend	الواجهة الأمامية	الوصول مباشرة من المستخدم والسماح له بالوصول إلى مزيد من الأجهزة أو البرامج.
F-Score	مقياس دقة الاختبار	هو مقياس لدقة الاختبار.حيث أن النتيجة F1 هي المعدل التوافقي للدقة والتذكر، حيث تصل درجة F1 إلى أفضل قيمة لها عند المستوى ١ (الدقة الكاملة والتذكير) والأسوأ عند ٠
Fully Voweled	مشكولة كلياً	جميع حروف الكلمة تحمل علامات التشكيل (الحركات أو الأحرف)
functional Structure (F Structure)	البنية الوظيفية	الهيكل الوظيفي هو واحد من الهياكل التنظيمية الأكثر شيوعا. وبموجب هذا الهيكل، يتم الاختيار وفقا لمجموعة متخصصة أو مماثلة من الأدوار أو المهام.
Functional Words	الكلمات الوظيفية	هي الكلمات التي ليس لها معنى معجمي يذكر أو لها معنى غامض وتعبر عن العلاقات النحوية بين الكلمات الأخرى داخل جملة، أو تحديد موقف أو مزاج المتكلم.
Gazatteers	معجم جغرافي	المعجم الجغرافي هو قاموس جغرافي أو دليل يستخدم بالاقتران مع خريطة أو أطلس. ويحتوي عادة على معلومات تتعلق بالتركيب الجغرافي والإحصاءات الاجتماعية والسمات المادية لبلد أو منطقة أو قارة.

General Ontology for Linguistic Description (Gold)	الأنطولوجيا العامة لوصف اللغويات (غولد).	هو اتجاه في دراسة اللغة يقوم على الوصف، ويقدم وصفا موضوعيًا لأهم الفئات والعلاقات المستخدمة في اللغة الإنسانية.
Generic Letters	الأصوات الأساسية	هي الحروف الثمانية والعشرون وحروف المد الثلاثة التي تعتمد على مخرج محقق أو مقدر على خلاف الحروف الفرعية التي تخرج من مخرجين أو تتردد بين حرفين.
Genitive	مجرور ، مضاف	حالة نحوية تصف إلصاق كلمة بأخرى تحل منها محل التنوين لتشكل تركيبًا خاصًا يعرف بالتركيب الإضافي.
Genres	أنواع	النوع إحدى الكليات الخمس التي تشير إلى مجموعة من السمات التي تجمع أشياء محددة تتفق بهذه السمات، تستخدم في الأدب مثلاً سواء كانت مكتوبًا أو منطوقًا.
Genre-Specific Features	السمات المستندة إلى النوع	الاستفادة من السمات اللغوية في الوصف لتقييم أهمية الأجزاء الفردية أثناء إعداد الملخصات للنصوص.
Glides	همزة على نبرة	الهمزة التي تأتي على نبر في حال جاءت مكسورة، بعد كسر، بعد ياء مد.
Glosses	حواشي	ترجمة توضيحية أو تفسير لكلمة أو عبارة وتظهر في أسفل الصفحة عادة.
Glottal Stop	همزة القطع	الهمزة التي تكتب وتلفظ سواء كانت في بدء الكلمة أم سطه أم آخره.
Grammar Analysis	تحليل قواعدي	تحليل قواعدي لبنية الجملة
Grammar Rules	القواعد النحوية	هي قواعد بناء الجمل.
Grammatical	نحوي	سمة الكلام الملتزم بالنظام النحوي في اللغة.
Grammatical Structure	التركيب النحوي	هو تركيب الكلمات والعبارات في الجملة.
Graph Theory	نظرية الرسم البياني	هو دراسة الرسوم البيانية، وهي الهياكل الرياضية المستخدمة لنمذجة العلاقات الثنائية بين العناصر.

Grapheme Segmentation	تقطيع الحروف	تقسيم تسلسل حروف أي كلمة في أي لغة في سلسلة من أصغر وحدة من نظام الكتابة الخاص بتلك اللغة سماعيا.
Graphemes	حروف	محاكاة كتابية على شكل رسم لأصغر وحدة صوتية لها وظيفية تمييزية وليس لها وظيفة دلالية في أي لغة.
Gray Level Transformation	معادلة شدة الإضاءة	يتميز المظهر المرئي للصورة بشكل عام بخاصيتين: السطوع والتباين. يشير السطوع إلى مستوى الشدة الكلي، وبالتالي يتأثر بقيم مستوى (كثافة) الرمادي الفردية لكل (بكسل) داخل الصورة.
Guesser	مخمن	برنامج يقوم بإنتاج أوصاف صرفية دلالية لأشكال الكلمات غير المعروفة للمحلل الصرفي.
Guidelines	المبادئ التوجيهية	هو بيان يحدد مسار العمل، ويهدف المبدأ التوجيهي إلى تبسيط عمليات معينة وفقا لممارسة روتينية أو ممارسة سليمة.
Half Vowel–Consonants	أشباه حروف العلة (في اللغة الإنجليزيةw and y)	هو الصوت الذي يشبه صوتيا حرف العلة، ولكن يعمل كحد للمقطع، وليس كنواة مقطع. أمثلة من الأحرف شبه في اللغة الإنجليزية هي الحروف الساكنة y وw ،
Handwriting	الخط أو الكتابة اليدوية	هو الكتابة التي تتم باليد وفق قواعد الكتابة والإملاء في لغة ما.
Handwriting Recognition Corpora	ذخيرة لغوية لتمييز الكتابة اليدوية	هي مجموعة كبيرة ومنظمة من النصوص المكتوبة يدويا يستطيع جهاز الحاسوب استلامها وتفسير الإدخال المكتوب بخط اليد مثل الوثائق الورقية.
Handwriting Recognition System	نظام التعرف إلى الكتابة اليدوية	هو نظام التعرف إلى الكتابة اليدوية.
Handwritten	مكتوب باليد	نص مكتوب باليد وليس مطبوعاً.

Handwritten Text	النص المكتوب بخط اليد	هي النصوص التي تمت كتابتها بخط اليد.
Hearst's Algorithm	خوارزمية هيرست	تطبيق الأنماط النحوية المعجمية لإظهار ارتباط مجموعات كبيرة من النصوص ببعضها بدقة عالية، بما في ذلك التطبيق المبكر لها إلى شبكة الكلمات وتستخدم هذه الخوارزمية على نطاق واسع في تطبيقات النصوص.
Heterogeneous Vocabulary	مفردات متعددة المصادر	قائمة مفردات ذات أصول لغوية مختلفة
Heuristics	المقاربة العملية	هو أي نهج لحل المشكلة، والتعلم، أو الاكتشاف الذي يستخدم طريقة عملية لا يضمن أن يكون الأمثل، ولكن كافية للأهداف الفورية.
Hidden Markov Models (HMMs)	نموذج ماركوف المخفية	نموذج ماركوف المخفي هو نموذج إحصائي يفترض أن النظام الذي يتم نمذجته هو عملية ماركوف مع حالات غير مرصودة (أي مخفية).
Hierarchical Structure	الهيكل الهرمي	هو التنظيم الهرمي أو الهيكل التنظيمي حيث كل كيان في المنظمة، باستثناء الرأس، يخضع لكيان واحد آخر. هذا الترتيب هو شكل من أشكال التسلسل الهرمي.
Hierarchy Cluster Technique	تقنية التصنيف الهرمي	هو طريقة للتحليل العنقودي الذي يسعى إلى بناء تسلسل هرمي للمجموعات.
Highly Inflected Language	لغة ذات خصائص تصريفية كبيرة	لغة غنية بالاشتقاقات.
Homographic	متجانس	كلمات مختلفة متطابقة في اللفظ ومختلفة في المعنى، وهي معكوس "مترادف"

أسلوب يستخدم للتعرف إلى الكلمات المكتوبة بخط اليد بشكل خاص، يساعد في الكشف عن الأساس استنادا إلى رسم بياني أفقي للإسقاط، بحيث يُحسب أسلوب كثافة الرسم البياني الأفقي لكل كلمة منفردة، ثم تحسب قيمة الأساس للتعرف إلى المناطق المرشحة التي قد تحتوي على القيمة المكملة.	طريقة منحنى الإسقاط الأفقي	Horizontal Projection Histogram Method
هي العملية التي لا يمكن اتمامها بطريقة آلية بشكل كلي إلا بمساعدة البشر.	التدخل البشري	Human Intervention
هو حقل بيني من تخصصات متعددة تشمل معظم التخصصات الفرعية للسانيات، واللغويات الحاسوبية، ومعالجة اللغة الطبيعية، وعلوم الحاسب، والذكاء الاصطناعي، وعلم النفس، والفلسفة، والرياضيات، والإحصاء.	تقنية اللغة البشرية	Human Language Technology
نهج هجين يجمع بين طريقتين مختلفتين للاستفادة من إيجابيات كلا الطريقتين.	مقاربة هجينة	Hybrid Approach
هي الطريقة التي تربط طريقتين أو أكثر في نفس الوقت للاستفادة من جميع المزايا.	الطريقة الهجينة	Hybrid Method
هو نظام ديناميكي يعتمد على دمج أكثر من نظامين بهدف الاستفادة من ميزاتهما ورفع كفاءة النظام الجديد.	نظام هجين	Hybrid System
كلمة ذات معنى واسع تشكل رأس فئة أو حقل يأتي تحتها كلمات ذات معانٍ أكثر تحديدا.	الإطار الدلالي العام	Hypernym
علاقة دلالية بين كلمتين فأكثر تقوم على أساس منطقي مثل علاقة الجنس بالنوع، حيث يكون الجنس هو الأعلى يقع تحته النوع؛ مثلاً الغراب نوع من جنس الطيور.	علاقة العام-بالخاص	Hypernyms-Hyponyms

Hypertext Markup Language (HTML)	لغة ترميز النصوص المتشعبة	هي لغة الترميز القياسية لإنشاء صفحات الويب وتطبيقات الويب، ، وتشكل مع صفائح الأنماط المتتالية وجافا سكريبت أساسًا من تقنيات حجر الأساس للشبكة العالمية.
Hyponym	الإطار الدلالي المخصص	كلمة ذات معنى أكثر تخصيصًا من مصطلح عام أو متفوق تشترك معه في الحقل الدلالي.
IBSN	الرقم الدولي المعياري للكتاب	أحد أنظمة التقييس الدولية، أداة عصرية سهلة تُمكّن الباحث أو القارئ من التعرف إلى أحد العناوين أو الطبعات الصادرة عن ناشر معين في بلد معين. وهو رقم فريد للعنوان أو للطبعة الواحدة.
Idioms	تعبير اصطلاحي	مجموعة من الكلمات يكون لها معنى محدد لا يعرف من المعنى الحرفي للكلمات الموجودة في التعبير أو التركيب.
Idiosyncratic	فرادي	الصفات الفريدة التي لا تحددها القواعد الصوتية للغة.
Imperative	صيغة الأمر	جملة تعبر عن الطلب بصيغة الأمر مثل اكتب قصيدة.
Indexing	الفهرسة	عملية جمع وتوزيع وتخزين البيانات لتسهيل استرجاع المعلومات بسرعة ودقة. يتطلب تصميم الفهرس مفاهيم من تخصصات متعددة: اللغويات، وعلم النفس المعرفي، والرياضيات، والمعلوماتية، وعلوم الحاسوب...
Indicative	تعييني	إشارة تحدد أو تعين دلالة كلمة ما لارتباطها بحالة معينة
Indispensable Sub-Task	مهمة فرعية أساسية	هي إحدى المهام الفرعية التي تندرج تحت مهمة رئيسية أخرى، لكن لا يمكن الاستغناء عنها.
Individual Cognitive Perception	الإدراك المعرفي الفردي	القدرة علي معالجة المعلومات من خلال الإدراك (المحفزات التي نتلقاها من خلال حواسنا المختلفة) ، والمعرفة المكتسبة من خلال الخبرة.
Infix	واسطة	اللواصق التي تأتي في وسط الكلمة
Infixes	الحُروف المَزيدة	هو إدراج حروف داخل الكلمة الجذعية.

Inflection	تصريف	تحويل الأصل الواحد إلى أمثلةٍ مختلفة لمعانٍ مقصودة لا تحصل إلا بها.
Inflectional and Cliticization Morphology System	نظام التوليد الصرفي	عبارة عن توليد التصريفات المختلفة التي يتم اشتقاقها من جذر الكلمة.
Inflectional Phrase	العبارة المصرفة	هي عبارة تحتوي على فئة مجردة، وهي الفئة الأساسية فيها وتحمل خصائص معينة.
Information Extraction	استخلاص المعلومات	هي مهمة استخراج المعلومات المنظمة تلقائيًا من وثائق غير منظمة و / أو شبه منظمة للقراءة الآلية.
Information Filtering	ترشيح المعلومات	هو نظام يزيل المعلومات الزائدة أو غير المرغوبة باستخدام أساليب شبه آلية أو محوسبة قبل تقديمها إلى المستخدم.
Information Retrieval	استرجاع المعلومات	هو نشاط استرداد المعلومات ذات الصلة من مجموعة من موارد المعلومات التي تم تخزينها سابقا.
Information Retrieval System	نظام لاسترجاع المعلومات	نظام يقوم بالحصول على موارد المعلومات ذات الصلة بالحاجة إلى المعلومات من مجموعة من الموارد.
Inputs and Outputs	مدخلات ومخرجات	هما عمليتان متعاكستان حيث إن المدخلات هي المعلومات التي تم إدخالها في نظام الحاسوب، ومن الأمثلة على ذلك: النص المكتوب، نقرات الماوس، إلخ. أما المخرجات هي المعلومات التي ينتجها نظام أو عملية من مدخلات محددة.
Intensifying Apposition	التوكيد اللفظي	نمط من أنماط التأكيد في اللغة، حيث تكرر اللفظة مرتين في الجملة للتأكيد.
Interlingua	لغة وسيطة	هي لغة مساعدة دولية مائلة، وضعت بين عامي ١٩٣٧ و ١٩٥١ من الرابطة الدولية للغة المساعدة، وتعد من أفضل المعاهدات التفاعلية الدولية الأكثر استخداما على نطاق واسع .

Interlingual Approach	مقاربة تعتمد لغة وسيطة	إحدى مقاربات الترجمة الآلية التي تعتمد لغة وسيطة ثالثة للترجمة بين لغتين
International Corpus of Arabic	الذخيرة العالمية للغة العربية	هي واحدة من الذخائر اللغوية للغة العربية.
Interoperability	قابلية العمل المشترك	هي قدرة نظم تكنولوجيا المعلومات المختلفة وتطبيقات البرمجيات على التواصل وتبادل البيانات والاستفادة منها.
Interpretation	تفسير	وهو نشاط يقوم على أساس شرح مباشر للنص بنفس لغته.
Interwiki Link	رابط قواعد بيانات متداخلة	هو رابط لإنشاء وصلات إلى العديد من قواعد البيانات على الشبكة العالمية.
Irregular Plural	جمع التكسير	هو جمع يتم بتغيير عدد من الأحرف في الاسم المُفرد ودون الاعتماد على قاعدة ثابتة.
Jaccard Coefficient	معامل جاكارد	كمية رقمية أو ثابتة وضعت لفحص تشابه العينات
Joint Rule-Based Model	النموذج المركب للقواعد	النمذجة الرياضية المركبة المبنية على القوانين أو القواعد
Jussive	المجزوم	حالة إعرابية تخص الفعل المضارع في اللغة العربية تظهر بقصر الحركة في آخر الفعل؛ فتكون بعلامة السكون أو استبدال حركة بحرف العلة المنتهي به الفعل؛ بحيث تكون الحركة من جنس الحرف.
Jussive Mood	صيغة الأمر	هو تركيب نحوي من الأفعال أو أسماء الأفعال لإصدار الأوامر.
Keyphrase	العبارة الرئيسية	هي العبارة المستخدمة في نص معين والتي تعبر عن مصطلح معين ولا يمكن الاستغناء عنها وغالبا تتكون من ثلاث كلمات أو أكثر.

Knowledge Management	إدارة المعرفة	إدارة ما يمتلكه الأفراد من مهارات تستند إلى المعرفة، وليس ما هو موثق في مستندات المؤسسة فقط. الهدف من إدراة المعرفة يرتبط بعملية اتخاذ القرار (Decision-making) في المؤسسات.
Knowledgebase	قاعدة معرفية	هي تقنية تستخدم لتخزين المعلومات المعقدة وغير المهيكلة المستخدمة من نظام الحاسوب.
Label	واصفة	سلسلة من الأحرف المستخدمة لوصف أو الإشارة إلى كائن أو كينونة معينة لتصنيفها في أحد البرامج.
Label Error Rate	نسبة أخطاء الوصف	نسبة حدوث خطأ في نظام الاكتشاف أو التنبؤ.
Language Model	نموذج اللغة	النماذج التي تهتم بالتعامل مع نظام البنية اللغوية.
Large Vocalized Language Models	النماذج الضخمة للغة المنطوقة	هي تلك النماذج المستخدمة للتعرف إلى الصوت والمعتمدة على بناء موارد تتضمن ملايين الكلمات.
Latent Semantic Analysis	آلية تحليل المفاهيم الكامنة في الوثائق	هي تقنية في معالجة اللغة الطبيعية، لا سيما الدلالات التوزيعية، وتحليل العلاقات بين مجموعة من الوثائق والشروط التي تحتوي عليها من خلال إنتاج مجموعة من المفاهيم المتعلقة بالوثائق والمصطلحات.
Learning-Based Approach	منهجية تعتمد التعلم الآلي	الاعتماد على التعلم الأساسي المبرمج مسبقاً للحاسوب لتنفيذ منهجية جديدة، حيث يعتمد المبدأ العام لهذه الأنظمة على تدريبه على جزء من البيانات المصنفة مسبقا، ثم فحص كفاءته على بقية البيانات.
Lemma	كلمة مجردة	الشكل المتعارف عليه أو نموذج التجريد لكلمة ما للتخلص من زوائدها.
Lemmatization	تجريد الكلمة من الزوائد	تجريد الكلمة من الزوائد بحيث تعود الكلمة إلى أصلها المعجمي بأقل عدد من الحروف.

Lemmatizer	المجذع	هو عملية تجميع أشكال مدمجة من كلمة بحيث يمكن تحليلها على أنها وحدة واحدة باستخدام قاعدة البيانات المعجمية والمحلل المورفولوجي المعدة مسبقًا.
Levantine Arabic	الشامية	اللغة العربية المتداولة في بلاد الشام (سوريا، ولبنان، والأردن، وفلسطين)
Levenshtein Edit Distance	مسافة ليفينشتين للاستبدال	هي إزالة أو إدراج أو استبدال حرف في سلسلة حروف الكلمة.
Lexeme-Based Morphology	التحليل الصرفي المعتمد على الجذع	هو التحليل الصرفي المعتمد على الجذع بحيث يتم استخراج الجذع Stem للكلمة المحللة فقط.
Lexica	معاجم	هي مجموع مفردات لغة أو عصر أو شخص أو فرع من المعرفة.
Lexical	مُعجمي	إشارة إلى الأشياء المتعلقة بمجموع كلمات لغة ما، أو معجم لغوي.
Lexical Ambiguity	غموض مُعجمي	هو مشكلة منتشرة في معالجة اللغة الطبيعية تربط بوجود تشابهات لفظية أو دلالية في كلمات معينة، ومع ذلك تتوفر معلومات كمية قليلة حول مدى المشكلة أو حول تأثيرها على نظم استرجاع المعلومات.
Lexical Collocations	التوافق المعجمي	هي عبارة عن كلمتين يتم دمجهما مع بعضهما للحصول على مصطلح معين. وعادة تكون هذه الكلمات (اسم، صفة، فعل أو ظرف).
Lexical Co-Occurrence	تلازم معجمي	هو مصطلح لغوي يمكن أن يعني التوافق بين كلمتين أو أكثر بحيث يردان في النص مع بعضهما في ترتيب معين.
Lexical Features	سمات مُعجمية	هي مجموعة من الخصائص تتعلق بكلمات أو مفردات اللغة.
Lexical Functional Grammar (LFG)	النحو المعجمي الوظيفي	هو إطار لقواعد اللغة أو النحو في اللغويات النظرية يعتمد على نحو العبارة كبديل من نحو المعتمد.

هو معيار المنظمة الدولية للتوحيد القياسي للمعالجة اللغوية الطبيعية ومعجم القاموس القابل للقراءة الآلية، ويهدف إلى توفير نموذج مشترك لإنشاء الموارد المعجمية واستخدامها، وإدارة تبادل البيانات فيما بين هذه الموارد.	إطار الترميز المعجمي	**Lexical Markup Framework**
هي قاعدة بيانات تتكون من قاموس واحد أو عدة قواميس.	الموارد المعجمية	**Lexical Resources**
هي عملية وضع حدود بين الكلمات أو المقاطع في اللغات الطبيعية سواء كانت بشرية أو اصطناعية.	تقطيع النص معجميا	**Lexical Segmentation**
هو حقل فرعي من علم الدلالة اللغوي. ويعنى بدراسة الدلالات المعجمية.	الدلالات المعجمية	**Lexical Semantics**
هو مجموعة من القواعد الأساسية التي تحدد كيفية كتابة البرامج في لغة بعينها، ويحدد هذا الهيكل أدنى مستوى لبناء الجملة والهيكل المستخدم في لغة البرمجة.	بنية معجمية	**Lexical Structure**
تصنيف وتحليل المواد المعجمية والاختلافات وأوجه التشابه في البنية الدلالية المعجمية لغويا وعلاقة المعنى المعجمى بمعنى الجملة وبناء الجملة.	الأنماط التركيبية المعجمية	**Lexical Syntactic Patterns**
هو لغوي ساهم في علم المعجم، فأوجد نظرية أو قام بجمع كلمات لغة ما وفق نظام خاص.	مؤلف معاجم (مُعجَميّ)	**Lexicographer**
هي قاعدة تقوم على بناء شبكة دلالية هائلة تربط المفاهيم والكيانات المسماة بمساعدة العلاقات الدلالية الكبيرة.	قاعدة معرفة معجمية	**Lexicographic Knowledge Base**
مادة لسانية تعنى بتأليف المعاجم والقواميس، وهي فرع من اللسانيات التطبيقية المعنية بتصميم وبناء معجم للاستخدام العملي.	معجمية	**Lexicography**

فرع من اللسانيات التطبيقية يعنى بتأليف المعاجم والقواميس، بجمع كلمات لغة ما وتعريفها وتصنيفها، وإيجازها في مقولات محددة على شكل عنوان جانبي. يهتم هذا الفرع أيضا بالشرح المصور، وبإعطاء أمثلة وتعبيرات للكلمة، لتوضيح المعاني والدلالات معا.	الصناعة المعجمية	Lexicography
هو فرع من اللسانيات يهتم بدراسة الكلمات مبنى ومعنى، والعلاقات الدلالية بينها، وسياقات استخدامها في لغة من اللغات.	علم المعاجم	Lexicology
كتاب يشتمل على عدد كبير من مفردات اللغة مرتبة ترتيبا معينا، مقرونة بطريقة نطقها وشرحها وتفسير معانيها وطرق استخدامها.	معجم	Lexicon
هو نموذج لنمذجة المعاجم والقواميس التي يمكن قراءتها آليا وترتبط بالويب الدلالي و مجموعة البيانات المرتبطة بها.	نموذج توصيف المعجم	Lexicon Model for Ontologies (lemon)
هي هياكل أو مخططات لغوية معممة تشير إلى العلاقات الدلالية بين المصطلحات، ويمكن تطبيقها على تحديد المفاهيم الرسمية والعلاقات المفاهيمية في نص اللغة الطبيعية.	أنماط دِلالية- مُعجميّة	Lexico-Syntactic Patterns
هو ربط الحروف بحيث ينضم اثنين أو أكثر من الرسوم البيانية أو الحروف بصورة واحدة.	تشابك الحروف	Ligature
التخلص من الزوائد القبلية والبعدية للكلمة.	التجذيع الخفيف	Light Stemmer
عملية التخلص من زوائد الكلمات القبلية والبعدية وإرجاعها إلى جذرها الأصلي.	تقليم الكلمة	Light Stemming

وجود فواصل موزعة خطيا	فواصل موزعة خطيا	Linear Spaced
الشخص الذي يقوم بدراسة لغة ما في ظواهرها ومستوياتها المتعددة.	لغوي	Linguist
هي عملية استخدام المتعلمين لنظام اللغة استخدامًا صحيحًا، بما في ذلك استخدام قواعد اللغة والنطق والمفردات.	الدقة اللغوية	Linguistic Accuracy
قاعدة لغوية لتشكيل الكلمات، أو تحويلها. وهي قاعدة تصف تحويل بنية نحوية واحدة إلى بنية نحوية أخرى ذات صلة.	قاعدة مؤسِسة للغة	Linguistic Based Rule
هي مجموعة من السمات التي تخضع لها كل لغة في ترتيب كلماتها وقواعدها، ويتم الالتزام بهذا الترتيب في تكوين الجمل والعبارات، فإذا اختل هذا النظام في ناحية من نواحيه لم يحقق الكلام الغرض منه وهو الإفهام، وهذا يعنى أن للغة خصائص محددة بها، وبالتالي لها استقلالها.	الخصائص اللغوية	Linguistic Features
هي طريقة لنشر البيانات المنظمة بحيث يمكن أن تكون مترابطة، وأن تصبح أكثر فائدة من خلال الاستعلامات الدلالية.	البيانات المرتبطة	Linked Data
هي سمة وصفية مرفقة بارتباط تشعبي للمستندات من أجل تعريف نوع الارتباط أو العلاقة بين المستند المصدر والوجهة. يمكن استخدام هذه السمة من الأنظمة المؤتمتة، أو يمكن تقديمها إلى المستخدمين بطريقة مختلفة.	علاقات الوثائق المرتبطة	Linked Documents Relationships
ترتبط البيانات المفتوحة المرتبطة بالبيانات التي يتم إصدارها بموجب ترخيص مفتوح، والتي يمكن إعادة استخدامها مجانا.	البيانات المفتوحة المرتبطة	Linked Open Data (LOD)
هي حالة نحوية تشير إلى موقع.	ظرف المكان	Locative Preposition

Long Short-Term Memory (LSTM)	ذاكرة قصيرة طويلة الأمد	وحدة بناء طبقات الشبكة العصبية المتكررة، وهي ذاكرة قصيرة لكنها طويلة الأمد.
Low Recall	استرجاع ضئيل	يطلق على النظام صفة الاسترجاع الضئيل أو القليل عندما تكون نسبة الاسترجاع لبيانات تطابق معايير معينة قليلة، وهو مؤشر سلبي على كفاءة النظام.
Machine Learning (ML)	تعلم الآلة	هو أحد فروع الذكاء الاصطناعي الذي يهتم بتصميم وتطوير خوارزميات وتقنيات تسمح للحواسيب بامتلاك خاصية "التعلم" دون أن تكون مبرمجة بشكل صريح.
Machine Readable	مقروء آليًا (الحاسوب)	هي البيانات الوصفية التي يمكن معالجتها وقراءتها بسهولة بواسطة جهاز حاسوب.
Machine Translation	ترجمة آلية	الترجمة من لغة إلى أخرى باستخدام الحاسوب
Machine Translation using Semantic Web (SWMT)	الترجمة الآلية باستخدام الشبكة الدلالية	وهي الترجمة الآلية التي تتم عن طريق الحاسوب مع الأخذ بعين الاعتبار المعاني الدلالية للكلمات والجمل للحصول على ترجمة ذات معنى وأكثر منطقية.
Machine-Aided Human Translation (MAHT)	ترجمة نصف آلية	هي نوع من أنواع الترجمة تعتمد على الإنسان بشكل خاص مع الاستعانة بالكمبيوتر في بعض الأحيان.
Machine-Readable Way	طريقة القراءة الآلية	هي الطريقة التي يمكن الاعتماد عليها في معالجة وقراءة البيانات الوصفية بواسطة جهاز الحاسوب
Manual Classification	تصنيف يدوي	بناء نماذج التصنيف من مجموعة بيانات المدخلات بشكل يدوي.
Mapping	عنونة	التخطيط المسبق لمشكلة عن طريق ترميزها برموز خاصة ذات دلالة.

Markov Clustering	عنقدة ماركوف	هي خوارزمية عنقودية غير خاضعة للرقابة سريعة وقابلة للتطوير للرسوم البيانية (تعرف أيضا باسم الشبكات)وتستند إلى محاكاة التدفق (العشوائي) في الرسوم البيانية.
Markov Model	نموذج ماركوف	هو نموذج عشوائي يستخدم لنموذج النظم المتغيرة عشوائيًا، ويفترض أن الحالات المستقبلية لا تعتمد إلا على الحالة الراهنة، وليس على الأحداث التي وقعت قبلها.
Masculine	ذكري	وجود صفات أو مظاهر تخص وترتبط تقليدياً بالرجال.
Maximal Marginal Relevance	الحدّ الأقصى الهامشي للمُلاءَمة أو الارتباط	هو مقياس لقياس مدى التباين بين البند قيد النظر وتلك العناصر المختارة بالفعل. فارتفاعها يعني أن العنصر الذي تم النظر فيه وثيق الصلة بالاستعلام ويحتوي على حد أدنى من التشابه مع العناصر المحددة السابقة.
Maximum Entropy	أقصى حالات الفوضى	هو حالة فيزيائية للنظام في أكبر اضطراب له، أو نموذج إحصائي من المعلومات الأقل ترميزا ، وتُعدُّ هذه الحالة من النظائر الهامة.
Media Monitoring Systems	أنظمة مراقبة الإعلام	هي الأنظمة التي تقوم برصد إنتاج المطبوعات، على الإنترنت وبث وسائل الإعلام. ويمكن إجراء ذلك لأسباب مختلفة، بما في ذلك السياسية والأمور التجارية والعلمية وغيرها.
Mel-Frequency Cepstral Coefficient (MFCC)	معامل التردد لتحليل الصوت	هي ميزة تستخدم على نطاق واسع في الكلام التلقائي والتعرف إلى مقاطع الصوت، وهي عبارة عن معاملات مجتمعة.
Meronym-Holonym	علاقة الجزء-بالكل	هي علاقة تربط جزء أو عضو في شيء ما بالجسم الكلي.
Metadata	البيانات الواصفة	هي البيانات التي تزود المستخدم بمعلومات عن جانب أو أكثر من جوانب البيانات.

Metaphors	الاستعارة	هي ادعاء معنى لفظة لغير ما وضعت له في أصل اللغة على سبيل الإعارة مع إيراد قرينة مانعة من إرادة المعنى عل حقيقته، بهدف تحقيق تأثير بلاغي ما.
Methodology	منهجية	هو التحليل النظري المنهجي للطرق المطبقة في مجال الدراسة
Middle Portion of Vowels	الجزء الأوسط من حروف العلة	هو صوت من أصوات حرف العلة e) ، وتستخدم في بعض اللغات المنطوقة ويرمز له بـ. □ə□
Minimally Supervised Methods	المناهج شبه الآلية	هي المنهجيات التي تتطلب تعلم كميات كبيرة من البيانات غير المعنونة، ولا تتطلب قواعد مصممة باليد أو التدريب اليدوي للبيانات.
Mnemonic	رابط التذكر	هي تقنية التعلم التي تساعد على الاحتفاظ بالمعلومات أو استرجاعها (تذكر) في الذاكرة البشرية.
Modelling	نمذجة	هي العملية التي تقوم على تمثيل نظام باستخدام القواعد والمفاهيم العامة.
Modern Standard Arabic	اللغة العربية الحديثة	اللغة العربية الرسمية المستخدمة في الإعلام والدوائر الحكومية (الخطاب الرسمي)
Modular Design	التصميم بالأجزاء	هو نهج التصميم الذي يقسم نظامًا إلى أجزاء أصغر تسمى وحدات، التي يمكن إنشاؤها بشكل مستقل، ومن ثم استخدامها في أنظمة مختلفة.
Module	وحدة برمجية	هي وحدة نمطية مكونة من برنامج أو جزء من برنامج يحتوي على واحد أو أكثر من الإجراءات، أو واحدة أو أكثر من وحدات تطويرها بشكل مستقل لتشكل برنامجا.
Monolingual	أحادي اللغة	هو مفهوم يشير إلى القدرة على التحدث بلغة واحدة فقط وفهمها والتعبير بها.

Montague Grammar	نحو (مونتاغيو)	هي نظرية قائمة على دلالات اللغة الطبيعية وعلاقتها ببناء الجملة، وهذا يعني أن معنى الكلمة بالكامل يمكن استخراجه من المعنى الكلي للجملة وطريقة تركيبها النحوي.
Mood	الحالة الإعرابية للفعل	الحالة الإعرابية للفعل في اللغة العربية، إما أن يأتي مرفوعًا أو منصوبًا أو مجزومًا.
Morpheme	وحدة صرفية	أصغر وحدة صوتية في اللغة لها معنى.
Morph-Lexical	معجم – صرفي	هو المعجم الذي يهتم بالتحليل اللغوي للكلمات اعتمادا على اشتقاقاتها الصرفية.
Morphological	صرفي	هو فرع اللغة الذي يتعامل مع شكل وهيكل الكلمة، ويهتم بدراسة ووصف الأنماط للكلمات اعتمادا على أنماط تشكيل الكلمة بلغة معينة، ويشتمل على التشكيل والاشتقاق والتصريف. ويعرف أيضا على أنه دراسة هيكل وتشكيل الكلمات، ووحدتها الأكثر أهمية هي المورفيم.
Morphological Analysis	تحليل صرفي	تحليل بنية الكلمة إلى مكوناتها الصرفية
Morphological Analysis Module	وحدة للتحليل الصرفي	وحدة لتحليل بنية الكلمات وأجزاء الكلمات
Morphological Analyzer	محلل صرفي	محلل لبنية الكلمات وأجزائها
Morphological Decomposition	التحليل الصرفي	هو تحليل الكلمة إلى مكوناتها الصرفية
Morphological Dependencies	الاعتماد الصرفي	وجوب توافق الكلمات في الجملة ببعض المزايا، مثل أن الفعل الذي يعقب الفاعل يجب أن يتوافق معه من ناحية التذكير والتأنيث؛ أي أن الفعل معتمد على الفاعل والفاعل معتمد على الفعل.

Morphological Derivation	اشتقاق	هي عملية تشكيل كلمة جديدة من كلمة موجودة، غالبا بإضافة بادئة أو لاحقة، على سبيل المثال كلمة قراءة شكلت من الأصل قرأ.
Morphological Disambiguation	فك اللبس الصرفي	فك اللبس الصرفي بالاستعانة بالسياق، ويتم ذلك اعتمادا على النظرية الاحتمالية للسياق المحلي للجملة. هو المهمة التي تهتم بتحديد التحليل الصرفي الصحيح للكلمة عن طريق إزالة التشويه الصرفي الذي ينتج من اشتقاقاتها المتنوعة.
Morphological Inflection	الانعطاف أو الانثناء الصرفي	هو عملية تغيير شكل الكلمة بحيث تعبر عن معلومات مثل عدد، شخص، حالة، الجنس...، مع المحافظة على الفئة النحوية للكلمة دون تغيير. على سبيل المثال، صيغة الجمع من الاسم في اللغة الإنجليزية وعادة ما تتكون من شكل المفرد عن طريق إضافة (s).
Morphological Patterns	الأنماط الصرفية	هي مجموعة من الأشكال أو العمليات التي تبني أشكالاً مختلفة من الوحدة الأساسية للكلمة، وتم تشكيل هذه الأنماط عن طريق الالتواء الصرفي أو التلصيق أو التضعيف أو الاشتقاق.
Morphological Quadruple	الرباعيات الصرفية	هي إحدى الميزات المستخرجة من الكلمات بهدف تسهيل التعامل معها
Morphological Segmentation	التجزئة الصرفية	تحليل هيكل الكلمات إلى أجزاء مثل الجذراً البادئات واللاحقات.
Morphological Tagging	وسم صرفي	إضافة حواشي لإعطاء معلومات صرفية عن النص
Morphological Weights	أوزان صرفية	أشكال صرفية تستخدم لتوزين الكلمات وتوصف بأنها قواعد دلالية لتشكيل مزيد من الكلمات من الجذر الأصلي ومثال ذلك : فعل فاعل مفعول يستفعل.

Morphologically-Rich Language (MRL)	لغة غنية صرفيا	لغة يغلب على مفرداتها القابلية للاشتقاق والتصريف.
Morphology	علم الصرف	العلم الذي يعرف به أحوال بنية الكلمة، وصرفها على وجوه شتى لمعان مختلفة
Morphology Ambiguity	الغموض الصرفي	هو عدم وضوح الكلمة من حيث وزنها وعدد حروفها وحركاتها وترتيبها، وما يعرض لذلك من تغيير وحذف، وما في حروف الكلمة من أصالة و زيادة.
Morpho-Syntactic Phenomena	الظواهر الصرفية النحوية	هي بعض التحليلات المتعلقة بأصول تكوين الجملة وقواعد الإعراب ومواضع الكلمات والخصائص التي تكتسبها الكلمة من ذلك الموضع، كالابتداء والفاعلية والمفعولية والتقديم والتأخير والإعراب والبناء أو ما يتعلق بتبيان وزن الكلمة وحركاتها وحروفها وترتيبها و ما يعرض لذلك كله.
Morphosyntactic Tagging	وسم الصّرف-نحوي	هي عملية وسم الكلمات حسب خصائصها النحوية والصرفية.
Morphotactics	التصريف	دراسة تكوين الكلمات من مكوناتها (الأجزاء الصرفية).
Multi-Document	مُتَعدّد-الوثائِق	هو محتوى ينطوي على معلومات تم الحصول عليها من أوراق متعددة او كتب متنوعة بخصوص موضوع ما.
Multilingual	متعدد اللغات	هي صفة تطلق على الأشخاص المتحدثين بأكثر من لغة أو النصوص المكتوبة بلغات متعددة
Multi-word Expresson (MWEs)	تعبير متعدد الكلمات	تعبير لغوي مكون من أكثر من كلمة، مثال: أبو أسعد
Mutation	تحول صوتي	تحول في حدة الصوت أو نبرته أو مخرجه لأسباب صوتية أو صرفية

N Grams	سلسلة محددة القيمة لعناصر لغوية	هو نهج يعتمد على مبدأ تقسيم النص إلى أقسام متتالية ومتداخلة حسب قيمة معينة يرمز لها بالرمز "N" فمثلا كلمة الطالب تصبح الطا لطا طال الب عند استخدام 3gram
Name Entity Phrase	عبارة دالة على كيان	عبارة مركبة من أكثر من كلمة تشير إلى كيان محدد
Named Entities	الأشياء المسماة	هو أي كائن فريد، يمكن تعريفه وفصله. وهو يشير إلى الأفراد أو المنظمات أو النظم أو وحدة من البيانات أو حتى مكونات النظام المتميزة التي تُعدُّ هامة في حد ذاتها.
Named entity Classification (NEC).	تصنيف أسماء الأشياء	هو تصنيف أسماء الأشياء التي يتم اكتشافها إلى مجموعة محددة من الفئات، فمثلا أسماء الأعلام التي تدل على مكان في فئة المكان وهكذا...
Named Entity Detection (NED)	الكشف عن أسماء الأشياء	هو اكتشاف وتحديد أسماء الأشياء من نص معين
Named Entity Recognition	تمييز أسماء الأشياء	قدرة الحاسوب على تمييز أسماء الأشياء بالاعتماد على الذكاء الاصطناعي، وهي التقنيات التي تهدف إلى التعرف إلى أسماء الأعلام وتمييزها
Natural Language	لغة طبيعية	اللغة البشرية التي يمكن للأطفال اكتسابها من أبائهم أو مربييهم بشكل عفوي دون تعليم أو إرشاد وأن يتعامل معها الناس كلغة أم ويطلق عليها حينئذٍمصطلح "لغة حية".
Natural Language Analyzer	محلل اللغة الطبيعية	هو أحد الأدوات المستخدمة في تحليل اللغة التي تطورت بشكل طبيعي عند البشر من خلال الاستخدام والتكرار دون تخطيط واعٍ ودراسة لأجزائها طبيعتها وفهمها وتفسيرها.

معالجة اللغات الطبيعية في مجال علوم الحاسوب واللغويات المعنية بالتفاعلات بين الحاسوب واللغات الطبيعية. وقد بدأت كفرع من الذكاء الاصطناعي التي تفرعت بدورها من المعلوماتية	حوسبة اللغات الطبيعية	Natural Language Processing (NLP)
هي بعض التطبيقات التي يمكنها الجمع بين الذكاء الاصطناعي واللغويات الحاسوبية وعلوم الحاسوب ودمجها باللغة البشرية الطبيعية لا سيما كيفية برمجة أجهزة الحاسوب لمعالجة مثمرة لكميات كبيرة من بيانات اللغة الطبيعية.	تطبيقات معالجة اللغة الطبيعية	Natural Language Processing Applications
هو مجموعة من الأنظمة تقدم بعض العناصر في الشاشة التي تسمح بالانتقال من صفحة إلى أخرى داخل الموقع الإلكتروني. وهي تطبيقات برمجية لاسترجاع المعلومات وعرضها على المستخدم من خلال الشبكة العالمية.	نظام التصفح	Navigation System
النفي القائم على المنطق	نفي	Negate
محاكاة الشبكة هي تقنية يقوم فيها برنامج برمجي بتبادل سلوك الشبكة من خلال حساب التفاعل بين مختلف كيانات الشبكة (أجهزة التوجيه ، والمفاتيح ، والعقد ، ونقاط الوصول ، والوصلات وما إلى ذلك). تستخدم معظم المحاكيات محاكاة أحداث منفصلة - نمذجة النظم التي تتغير.	شبكة المحاكاة والاختبار	Network Simulation and Testing
هو نهج للترجمة الآلية يستخدم شبكة عصبية اصطناعية كبيرة للتنبؤ باحتمال وجود سلسلة من الكلمات، ويعتمد عادة على نمذجة الجمل بأكملها في نموذج متكامل واحد.	الترجمة الآلية العصبية	Neural Machine Translation (NMT)
نظام محوسب يحاكي في تصميمه الدماغ البشري والجهاز العصبي ويستخدم لبناء أنظمة الذكاء الاصطناعي.	شبكة عصبية	Neural Network

Neural Network Techniques	تقنيات الشبكة العصبية	هي أنظمة الحوسبة المستوحاة بشكل غامض من الشبكات العصبية البيولوجية التي تشكل أدمغة الكائنات الحية. هذه الأنظمة "تتعلم" المهام من خلال النظر في الأمثلة بحيث تتحسن تدريجيًا في الأداء بدون برمجة خاصة بهذه المهام.
Nodes	عُقد	هي عناصر أو مكونات مرتبطة مع غيرها بعلاقات متعددة
Noise Reduction and Silence Removal	الحد من الضوضاء وإزالة الصمت	هو عملية إزالة الضوضاء من إشارة في جميع أجهزة التسجيل، سواء التناظرية أو الرقمية.
Nominal	اسمي	ما يتعلق وجوده بالاسم فقط.
Nominal and Verbal Clauses	الجمل الاسمية والفعلية	الجملة الاسمية هي ما تقدم فيها العنصر الاسمي، ويتكون تركيبها الأساسي من جزأين هما: المبتدأ والخبر. أما الجملة الفعلية فهي الجملة التي تبدأ بالفعل بأحد أنواعه الثلاثة الماضي والمضارع والأمر. وعادة ما تتكون من فعل فاعل على الأقل.
Nominal Sentence	الجُملة الاسمية	الجملة الاسمية هي ما تقدم فيها العنصر الاسمي، ويتكون تركيبها الأساسي من جزأين هما: المبتدأ والخبر، أو المسند إليه والمسند. فالعلاقة بين عنصري الجملة الاسمية هي علاقة الإسناد، فالمبتدأ موضوع، والخبر حديث عن هذا الموضوع، والمبتدأ محكوم عليه والخبر محكوم به.
Nominative	المرفوع	حالة من حالات الإعراب خاصة بالأسماء والأفعال المضارعة في اللغة العربية تظهر بعلامات مخصوصة في نهاية الكلمة.
Nonconcatenative Morphology	الصرف غير الإلصاقي	خاصية صرفية للغة العربية تعمل على اشتقاق كلمات جديدة من الجذور والأوزان الصرفية، وفي هذه العملية يتم إضافة أحرف زائدة داخلية

Normalization	تطبيع	عملية توحيد صيغ إدخال وتخزين واستخراج الكلمات من الحاسوب
Normalized Lexicons	معاجم مبسطة	هي المعاجم التي يتم اعادة تمثيل كلماتها بشكل يضمن المزيد من التبسيط في التعامل معها
Noun Phrase	العبارة الاسمية	وهي عباره تبدأ باسم ويكون هذا الاسم هو الكلمة الرئيسية فيها.
Nuisances	مصادر الإزعاج	هي مصادر معينة سواء أشخاص أو أشياء تسبب الإزعاج.
Number	العدد	العدد للأسماء إما مفرد أو مثنى أو جمع
Numeral	رقمي	هو المحتوى الذي يتكون من عناصر رقمية فقط
Off-Line	غير متصل بالشبكة	حالة الاتصال بشبكة الانترنت ليست فعالة
Off-Line System	نظام غير متصل	نظام غير متصل بالشبكة أو لا يتم تحديث بياناته في الزمن الحقيقي.
On-Line	متصل بالشبكة	حالة الاتصال بشبكة الانترنت فعالة
Online Recognition	التعرف المباشر	هي مجموعة من الأدوات البرمجية التي تستطيع التعرف إلى النصوص أو الصور مباشرة.
Ontological Relations	العلاقات الأنطولوجية	هي العلاقات التي تحدد طبيعة الوجود أو الواقع، بالإضافة إلى الفئات الأساسية للعلاقات. غالبًا ما تتعامل هذه العلاقات مع الأسئلة المتعلقة بالكيانات الموجودة أو ما يقال إنها موجودة ، وكيف يمكن تصنيف هذه الكيانات، المرتبطة ضمن التسلسل الهرمي، وتقسيمها.
Ontology	أنطولوجيا	هو مفهوم قد بدأ بالظهور من مبادرات الويب الدلالي المتعددة. ويتعلق بالتوصيف والتعرف إلى دلالات الكلام.

Ontology-Based Semantic Context Framework (OBSC)	إطار السياق الدلالي القائم على الأنطولوجيا	هو نموذج المحتوى الدلالي الذي يعتمد على الأنطولوجيا في بناء الحقائق
Opinion Mining	التّنقيب أو البحث عن الرأي	هي عملية البحث أو التنقيب عن الآراء باستخدام معالجة اللغات الطبيعية، وتحليل النصوص، واللغويات الحاسوبية.
Optical Character Recognition OCR	التعرف الضوئي على الحروف	تجري عمليّة التعرّف الضّوئيّ على نصّ محوْسب في صيغة الوثيقة المتبادلة (امتداد (pdf. أو بعد المسح الضّوئيّ (scanning) لإدخال ذلك النّصّ. وبعد أن يقرأ البرنامج النّص، يحلّل أشكال الصّور ويتعرّف إلى الحروف تلقائيّا.
Optimized Platforms Architectures	التراكيب الأمثل أطرا	أنواع الأطر أو الأنظمة ذات التراكيب المختلفة التي تسمح بتقديم أفضل الحالات عند التنوع في تنظيم هذه التراكيب
Orthography	الإملاء	نظام الكتابة، وهو ميزة تتعلق بكيفية تعيين أصوات اللغة اعتمادا على برنامج نصي معين. وهو مجموعة من القواعد لكتابة لغة ما و يتضمن رسم الحروف، الشرطة، الكتابة بالأحرف الكبيرة، فواصل الكلمة، وعلامات الترقيم...
Out of Vocabulary	الكلمات غير المعروفة	هي كلمات غير معروفة تظهر في لغة الاختبار ولكن ليس في المفردات المعروفة. وعادة ما تكون كلمات المحتوى مهمة مثل الأسماء والمواقع التي تحتوي على معلومات حاسمة لنجاح العديد من مهام التعرف إلى الكلام.
Out of Vocabulary Words	الكلمات غير المدرجة في القواميس	هي كلمات غير معروفة لا تظهر في قوائم المفردات المعروفة وذلك لأنه من المستحيل إدراج جميع كلمات اللغة الطبيعية في قوائم ثابتة.

Overlapping	التداخل	مصطلح يشير إلى تشارك عنصرين مشتركين وتقاطعهما في صفات مشتركة
Overt Morpheme	مقطع صرفي صريح	هو مقطع أصغر وحدة نحوية في اللغة يمكن أن يعبر عن معنى كلمة بحد ذاتها ويمكن أن تعبر عن مقطع ليس ذا معنىً كما في لواحق الكلمة.
Paper Dictionary	قاموس ورقي	هو كتاب ورقي يعطي قائمة من كلمات لغة ما حسب الترتيب الأبجدي ويشرح ما يعنيه، أو يعطي كلمة تقابل كلمة أخرى في لغة أجنبية
Parse Trees	تحليل شجري	هو عبارة عن تمثيل تخطيطي للهيكل المحلل للجملةٍ والذي يظهر قواعد اللغة و جذور الكلمات وبنيتها النحوية بغض النظر عن السياق.
Parsing Expression Grammars	تحليل القواعد النحوية التعبيرية	هو نوع من التحليل الشكلي لقواعد لغة لوصفها صوريا بهدف التعرف إلى مكوناتها.
Part of Speech (POS)	أقسام الكلام	هو استخراج أقسام الكلام من النص (أقسام الكلام مثل فعل ماض، فعل مضارع، فعل أمر، صفة، اسم، حال ..إلخ)، فيتم تصنيف كل كلمة في النص إلى القسم الذي يمثلها
Part of Speech Tagger	برنامج وسم أقسام الكلام	هو النموذج الذي يقوم بمهمة تحديد نوع الكلمة حسب أقسام النص
Part of Speech Tagging	وسم أقسام الكلام	تصنيف الكلام إلى أنواعه صرفياً ونحوياً
Partially Voweled	مشكولة جزئياً	بعض حروف الكلمة التي تحمل التشكيل (أحرف العلة القصيرة).
Particles	أدوات	هي أجزاء الكلام الذي لا يمكن أن يكون له علامة إعراب، وتستخدم هذه الكلمات ككلمة وظيفية مرتبطة بكلمة أو عبارة أخرى لنقل المعنى.

هو توصيف لكل قسم من أقسام النص حسب نوعه فمثلا نشير لكل قسم على أنه فعل ماض، فعل مضارع، فعل أمر، صفة، اسم، حال ..إلخ)، فتتم الإشارة إلى صنف كل كلمة في النص حسب القسم الذي يمثلها.	وسم نوع الكلمة	**Part-of-Speech Tagging**
هو اسم مشتق من حروف الفعل المتصرف المبني للمجهول؛ ليدل على من وقع عليه فعل الفاعل . مثل : سرق : مَسروق – استعمل : مُستعمَل. وهو شكل من أشكال الفعل في بعض اللغات يمكن أن يعمل بشكل مستقل كصفة.	اسم المفعول	**Passive Participle**
هو فعل لم يُعلم فاعله، أي مجهول الفاعل. وينوب عنه عادة المفعول به ويكون مرفوعا ويعرب نائب فاعل.	المبني للمجهول	**Passivization**
هو أساس من أساسات علم الصرف. وهو طريقة لوزن الكلمات في اللغة العربية، والتأكد من أنها تقع ضمن وزن معين لتكون وزنًا لأصول الكلمات يعبلا عنها بصيغة (فعل). حيث إن الفاء تقابل الحرف الأول، والعين تقابل الحرف الثاني، واللام تقابل الحرف الثالث مع الأخذ بعين الاعتبار علامات التشكيل.	وزن صرفي	**Pattern**
هي العملية التي يتم من خلالها استخراج وزن الكلمة بناء على جذرها وعلامات التشكيل. والوزن الذي تقوم عليها جميع الكلمات هو الفعل (فعل) وأوزانه.	عملية استخراج الوزن	**Pattern Extractor Process**
هي إيجاد الوزن الذي يناسب كل كلمة في الصرف أو مطابقة صورة مع أخرى في التعرف إلى الحروف أو الأصوات.	مطابقة الهيئة	**Pattern Matching**
هو إجراء إحصائي يستخدم لتحليل مجموعة من الملاحظات من المتغيرات المحتملة المترابطة إلى مجموعة من القيم للمتغيرات غير المترابطة خطيا و تسمى المكونات الرئيسية.	تحليل المكونات الرئيسية	**PCA Principal Component Analysis**

هو مقياس للارتباط الخطي بين متغيرين X و .Y وله قيمة بين +١ و -١ حيث ١ هو مجموع الارتباط الطولي الموجب، ٠ ليس هناك ارتباط خطي، و -١ هو مجموع الارتباط الخطي السلبي.	معامل ارتباط بيرسون	**Pearson Correlation Coefficient**
هو كل ما يتعلق، أو يلائم المعلم أو العملية التعليمية من حيث الأساليب التربوية والمخاوف التربوية وغيرها.	تعليمي	**Pedagogical**
إسناد الفعل إلى المتكلم أو المخاطب أو الغائب (أحد الخصائص الصرفية للأفعال والضمائر)	الإسناد	**Person**
هو عبارة عن طريقة لاستخراج جذر الكلمة غير الأصلي أو غير الحقيقي باستخلاص أفضل تعبير يعبر عن الكلمة	جذع وهمي	**Phantom Stem**
الوحدة الصوتية التي تتكون منها الكلمة المنطوقة	مقطع صوتي	**Phoneme**
هو عبارة عن مزيج من صوتين من أصوات حروف العلة بحيث يتم دمجهما ويلفظا كصوت واحد.	الجمع بين الأصوات	**Phonetic Combination**
هي مجموعة من القواعد المتعلقة بالعملية الصوتية ولفظ الكلمات بشكل صحيح.	قواعد صوتية	**Phonetic Rules**
هي مجموعة من الخصائص المكونة للبنية الصوتية التي يمكن تحليلها طبقا للنظرية الصوتية.	الخصائص الصوتية	**Phonological Properties**
هو علم من علوم الصوتيات يشكل نظام العلاقات التبادلية بين الكلام (الأصوات) التي تشكل المكونات الأساسية للغة مثلاً المقاطعاً التشديداً اللهجةاً النغمة وغيرها.	علم الأصوات	**Phonology**
هي عبارة أو مجموعة من الكلمات التي تحمل في الغالب معنى اصطلاحيًا خاصًا ، وهذا المعنى هو مرادف تقريبا للتعبير. والعبارة هنا هي مجموعة من الكلمات (أو ربما كلمة واحدة) تعمل كمكون في تركيب الجملة.	العقد الجملية	**Phrasal Nodes**

Phrase Chunking	تقطيع العبارات	تقطيع الجملة إلى مكوناتها من الجمل الأصغر وأشباه الجمل
Phrase Structure Grammars	نحو بناء العبارة	هي تلك القواعد النحوية التي تقوم على علاقات مستقلة بدلا من العلاقات التبعية المرتبطة بالاعتماد على القواعد النحوية. أي عدد من النظريات ذات الصلة لتحليل اللغة الطبيعية تكون مؤهلة كمجموعة قواعد.
Phrase Structure Theory	نظرية نحو العبارة	هي نظرية تقوم على قواعد بنية العبارة و إعادة الكتابة المستخدمة لوصف بنية لغة معينة وترتبط ارتباطا وثيقا بالمراحل المبكرة من القواعد.
Phrase-Based	معتمد على العبارة	هي النهج الذي يعتمد على العبارات أو الجمل وتحليلها لاستخراج المعلومات والمعرفة
Plural	الجمع	هو مفهوم يدل على الكثرة أو يدل على مجموعة مكونة من عنصرين أو أكثر.
Polarity	قطبية الرأي	تحديد القصد من وراء الرأي، فمثلا يمكن أن يكون الهدف من الرأي الإيجابية أو السلبية أو أن يكون حياديًا
Possessive	الملكية (تملك)	هي كلمة أو بناء نحوي يستخدم للإشارة إلى علاقة حيازة بمعنى واسع.
Possessive Pronouns	ضمائر الملكية	هو الضمير الذي يستخدم للإشارة إلى شيء من نوع معين ينتمي إلى شخص ما ، كما في "هل يمكنني استعارة قلمك؟ وبالمثل أيضا كلمة "ملكك" أو "ملكهم"
Post –Editing	بعد التعديل- التحرير	هو خارج عملية التعديل أو ما يتبع هذه العملية
Post-Processing	المعالجة اللاحقة	هي عملية المعالجة الثانوية التي تستخدم لتصفية البيانات والتخلص من الضوضاء التي لم يتم التعرف إليها في مرحلة المعالجة الأولية

Potentially Ambiguous Noun Phrases (NPs)	عبارات الأسماء محتملة الغموض	هي الأسماء التي تحمل نوعا من الغموض في دلالتها فمثلا قد يكتب المستخدم في محرك البحث : "كم عمر باراك أوباما"، "ومن هو متزوج من"،فهنا الضمير "هو" غير معروف على من يعود ، وهناك أنظمة طورت للاستعاضة عن هذا الضمير بالاسم الذي يسبقه للتقليل من الغموض
Pragmatics	برغماتية	دراسة العلاقات بين رموز اللغة، ومعانيها وفق حالات استخدام اللغة وسياقاتها.
Precision	الضبط	هو أسلوب من أساليب تقييم دقة النظام، والدقة تشير إلى القرب من اثنين أو أكثر من القياسات لبعضها بعضًا.
Predefined Classes	بنى معطيات معرفة مسبقاً	هي الفئات التي تحمل مسمى محددًا ومعرفًا مسبقا بواسطة الخبراء
Pre-Emphasis	التهيئة المسبقة	يشير هذا المصطلح إلى التهيئة المسبقة التي تكون مصممة لزيادة في حجم بعض الترددات (نطاق التردد) مقارنة بحجم الترددات الأخرى (الأقل عادة) من أجل تحسين المجموع الكلي لنسبة الإشارة إلى الضوضاء عن طريق التقليل إلى أدنى حد من الآثار السلبية لهذه الظواهر المسببة للتشوه.
Prefix	سابقة	السوابق هي الزوائد التي تتم إضافتها في أول الكلمة.
Preposition	حرف الجر	أحد أنواع الكلم الثلاثة (اسم، فعل، حرف) لا يقوم معناه في ذاته، مثل: من، على، إلى ...
Prepositional Phrase	شبه الجملة	تركيب لغوي لا يقوم بذاته، مكون عادة من حرف جر واسم يتلوه، أو ظرف يضاف إلى اسم آخر.
Preprocessing	معالجة مسبقة	الخطوات التحضيرية للبيانات قبل المعالجة، وتشمل حذف الزوائد والأرقام والكلمات اللاتينية وغيرها.

توزيع الاحتمال هو دالة حسابية يتم التعبير عنها بعبارات بسيطة ، ويمكن الاعتماد عليها لتوفير احتمالات حدوث نتائج مختلفة في التجربة. على سبيل المثال ، إذا تم استخدام المتغير العشوائي X للإشارة إلى نتيجة رمي العملة ("التجربة") ، فإن توزيع الاحتمالية لـ X سيأخذ مجموع الاحتمالات الرقمية المحتملة الحدوث.	توزيع الاحتمالية	Probability Distribution
الزوائد المتصلة في بداية الكلمة كحروف العطف وحروف الجر والنداء وأدوات التعريف.	زائدة أولية	Proclitic
الضمير (اختصاره (pro هي كلمة يستعاض بها عن عبارة اسمية أو عن اسم صريح. وهي حالة خاصة من إعادة التشكيل.	ضمير	Pronoun
القدرة على النطق أو ما تكون قابلة أن تلفظ وتعرف أيضا بأنها "مجموعة حروف منطوقة"	مهجأ	Pronounce-able
هو معجم مصمم لتمكين الجمل القابلة للنطق والقابلة للتشغيل من الانطلاق بشكل مسموع بحيث يتم التعرّف إلى الكلام ومحركات تركيب الكلام ضمن تطبيقات التصفح الصوتي. وتهدف هذه المعجمات إلى تسهيل الاستخدام من المطورين بينما تدعم المواصفات الدقيقة لمعلومات النطق.	معجم النطق	Pronunciation Lexicon
هو عبارة عن عملية التشييك ومتابعة ظهور أخطاء في نص مكتوب وتصويبه	التدقيق اللغوي	Proofreading
هي عبارة عن أسماء العلم التي تدل على اسم شخص أو إله أو مكان محدد أو اسم شركة محددة	أسماء الأعلام	Proper Nouns
هو نموذج أو نهج تصميمي يهدف إلى بناء أولي لحل مشكلة معينة.	نموذج مبدئي	Prototype

Psycholinguistics	علم النفس اللغوي	هو دراسة العوامل النفسية والعصبية التي تمكن البشر من اكتساب واستخدام وفهم وإنتاج اللغة. يهتم النظام الأساسي بالآليات التي تتم فيها معالجة اللغات وتمثيلها في الدماغ.
Punctuation	علامات الترقيم	إن علامات الترقيم (التي كانت تُسمى أحيانًا بالإشارة) هي استخدام التباعد، والعلامات التقليدية ، وبعض الأجهزة المطبعية كوسائل مساعدة للفهم والقراءة الصحيحة ، بصمت أو بصوت عال ، للنصوص المطبوعة والمكتوبة بخط اليد.
Python	لغة البرمجة بايثون	هي واحدة من أشهر لغات البرمجة التي تدعم معالجة اللغات الطبيعية عن طريق توفير حجم مهول من المكتبات لمعالجة المشكلات اللغوية
Quad-Grams	تسلسل من أربع كلمات	هو تجزئة المحتوى إلى مجموعة مقاطع بحيث يتكون كل مقطع من أربعة عناصر مع السماح بتداخل أي مقطع بآخر بحيث يبدأ من العنصر الثاني وهكذا
Qualitative	النوعي	هو النهج الذي يستند إلى القيمة النوعية أو المقدار النوعي
Quantitative	الكمي	هو النهج الذي يعتمد على القيمة الكمية
Query	استعلام	هي البحث والتفتيش عن معلومة معينة او عن الحقائق المرتبطة بموضوع معين
Query Expansion (QE)	توسيع البحث أو الاستعلام	هو عملية زيادة حيز الاستعلام عن طريق الاستفادة من العلاقات المرتبطة بالعنصر المراد البحث عنه
Question Answering QA	سؤال جواب	هي تقنية تهدف إلى إيجاد أقرب إجابة لأي سؤال ومن الأمثلة عليها تقنيات البحث عن الأسئلة على محركات البحث

يقصد بهذا المصطلح التعبير الجذري أو أصل الشي وفي معالجة اللغات الطبيعية فإن هذا المصطلح يعبر عن جذر الكلمة	أساسي	**Radical**
هو تنظيم البيانات بطريقة تسهل عرضها وتبسط فهمها والوصول إليها حسب عدة معايير كالاسم أو الحجم أو اللون أو غير ذلك.	الترتيب	**Ranking**
هي المصطلحات الفريدة وقليلة الحدوث وتُعدُّ تميزية للدلالة على المحتوى الذي تكون فيه نظرا لقلة ظهورها	مصطلحات نادرة	**Rare Terms**
مستودع يجمع كمًا من النصوص التي يتم التقاطها من بيئاتها كما هي دون أي تنقيح	ذخيرة لغوية خام	**Raw text Corpora**
محتوى قابل للقراءة وسهل التعرف على دلالته	مقروء أو قابل للقراءة	**Readable**
هو مقياس للحكم على فعالية نظام معين للمعالجة الآلية للنصوص، ويسمى الاستدعاء (المعروف أيضا بالحساسية) هو جزء من الوثائق ذات الصلة التي استردت من العدد الإجمالي للوثائق ذات الصلة.	مسترجع	**Recall**
هي تقنية طورت لتكيف النطاق لاستخدامه في التعرف إلى الكيان المسمى.	التعلم بالاستدعاء	**Recall-Oriented Learning**
تكرار أو سرد كلام معين أو التكلم بصوت عالٍ (شيء تم حفظه أو التدرب عليه)	يتلو	**Recite**
هي التقنيات التي تهتم بالتعرف إلى نمط معين بين عدة أنماط بناء على خصائصه	خوارزمية التعرف	**Recognition Algorithm**
نظام محوسب يحاكي في تصميمه الدماغ البشري والجهاز العصبي ويستخدم لبناء أنظمة الذكاء الاصطناعي من خلال طبقات متكررة من الخلايا العصبية الاصطناعية.	الشبكات العصبية المتكررة	**Recurrent Neural Networks**

Recursively	متعاقب	تعريف الشيء بدلالة ذاته أو نوعه. ويستخدم في مجموعة من التخصصات التي تتراوح بين علم اللغة والمنطق. وفي نظام الكتابة هو اتصال الحروف لتكوين الكلمات.
Redundancy Data	بيانات فائضة	تكرار البيانات هو وجود بيانات إضافية للبيانات الفعلية بما يسمح بتصحيح الأخطاء في البيانات المخزنة أو المرسلة. يمكن ببساطة أن تكون البيانات الإضافية نسخة كاملة من البيانات الفعلية ، أو اختيار أجزاء من البيانات التي تسمح باكتشاف الأخطاء وإعادة بناء البيانات فقط
Reference Arabic Dataset	مجموعة بيانات عربية مرجعية	هي البيانات العربية التي تحدد مجموعة القيم المسموح باستخدامها في حقول البيانات الأخرى.
Reference Pattern	نمط مرجعي	يفترض هذا النمط أن النموذج المرجعي لاي نظام يلبي المتطلبات الأساسية للاتجاه الثابت.
Regular Expression	تسلسل متكرر للحروف	صيغة جبرية قيمتها نمط يتكون من مجموعة من سلاسل الحروف.
Relevance Feedback	التغذية الراجعة للملاءمة	تعليقات الملاءمة تعد ميزة لمعظم أنظمة استرجاع المعلومات. وتكمن الفكرة وراء التعليقات ذات الصلة في تحقيق النتائج التي يتم إرجاعها في البداية من طلب بحث معين ، لجمع تعليقات المستخدمين ، واستخدام المعلومات حول ما إذا كانت هذه النتائج ذات صلة بتنفيذ طلب بحث.

Resource Description Framework (RDF)	إطار وصف الموارد	هي مجموعة من مواصفات اتحاد شبكة الويب العالمية (W3C)التي تم تصميمها في الأصل كنموذج بيانات وصفية. وقد أصبحت تستخدم كطريقة عامة للوصف المفاهيمي أو نمذجة المعلومات التي يتم تنفيذها في موارد الويب ، وذلك باستخدام مجموعة متنوعة من تدوينات الجملة وتنسيقاتها.
Results Clusters	مجموعات النتائج	التحليل العنقودي أو التجميع هو مهمة تجميع مجموعة من الكائنات بطريقة معينة بحيث تكون الكائنات في نفس المجموعة (تسمى الكتلة) أكثر تشابهاً (بمعنى ما) مع بعضها بعضًا عن تلك الموجودة في المجموعات الأخرى (المجموعات). إنها مهمة رئيسية للتنقيب الاستكشافي عن البيانات.
Retrieval System	نظام استرجاع	هو عملية الحصول على بيانات أو معلومات ذات صلة من مجموعة من موارد المعلومات. ويمكن أن تستند عمليات البحث إلى نص كامل أو فهرسة محتوى. ويعرف أيضا بأنه علم البحث عن المعلومات في مستند، والبحث عن المستندات نفسها.
Reusable Lexical Data Bases	قواعد البيانات المعجمية القابلة لإعادة الاستخدام	هي ملفات قاعدة بيانات معجمية قابلة لإعادة الاستخدام ، والتي تستند إلى العمل الميداني أو التي تصمم عن طريق تاريخ البحث ، والاستفادة من المطابقة للمعايير المعمول بها بحيث يربط بين جميع التحليلات اللغوية و علاقة الوثيقة بالتسجيلات الأساسية والمجموعة النصية
Reusable Linguistic Resources	الموارد اللغوية القابلة لإعادة الاستخدام	هي عبارة عن الموارد والأدوات اللغوية المتاحة والتي يمكن إعادة صياغتها واستخدامها

Reversible	قابل للانعكاس	هي المواد التي تقبل إعادة استخدامها وتوضيفها بشكل آخر وتكرار استغلالها
Rework	إعادة العمل	إعادة تنفيذ المهمة مرة أخرى
Root	الجذر	أصغر وحدة صوتية لها معنى وقابلة للتشكيل في صيغ صرفية أخرى، يجب ألا يقل عن ثلاثة أحرف و الجذر هو الحروف الأصلية في الكلمة.
Root Extraction	استخراج الجذور	يستخرج الجذر الحقيقي للكلمة بإزلة الزوائد على أصل الكلمة
Root-Pattern Methodology	الصرف المعتمد على جذر ووزن الكلمات	التحليل الصرفي الذي يعتمد على جذر و وزن الكلمة كأساس للتحليل.
Root-Pattern Schemes	خرائط الجذور والأوزان	العلاقات التي تربط أوزان الكلمات وجذورها على مستويي المبنى والمعنى.
Rule Based	قواعدي	استخدام القواعد والأسس للقيام بمهمة معينة
Rule-Based Approach	منهجية تعتمد القواعد	الاعتماد على قواعد تعطى للحاسوب لتنفيذ منهجية معينة
Rule-Based Method	منهج قواعدي	هو عبارة عن النموذج أو النهج الذي يطبق تبعا لقواعد وأسس محددة
Runtime	زمن التشغيل	هو الوقت المستغرق للقيام بأداء وتشغيل مهمة معينة
Search Engine	محرّك بحث	استخدام مجموعة متنوعة من الأطر ولغات البرمجة للتحليل والتوليد، بدرجات مختلفة من التطور والمتانة والكفاءة التي تم تصميمها للبحث عن معلومات عل شبكة الإنترنت العالمية.
Search Tool	أداة بحث	هي اداة التي تعني بمهمة البحث والتنقيب عن محتوى معين
Segmentation	تجزئة أو تقطيع الكلمة أو النص	مهمة تقسيم أو تجزئة مجموعة كبيرة إلى عدة مجموعات أصغر حسب معيار معين

Semantic	دلالي	هي دراسة لغوية وفلسفية للمعنى اللغوي، ولغات البرمجة، والمنطق الرسمي، والمفاهيم السيميائية وغيرها . وهي تهتم بالعلاقة بين الدلالات – مثل الكلمات والعبارات والعلامات والرموز – وما تمثله من دلالاتها.
Semantic Analysis	تحليل دلالي	تحليل النص للوصول إلى المعنى من خلال تحليل الروابط الدلالية بين الكلمات والعبارت والجمل والفقرات.
Semantic Dependencies	الاعتمادات الدلالية	هي العلاقات الدلالية التي يتم تحديدها عن طريق المستندات أو الافتراضات وحججها. ويمكن الاعتماد على العلاقات الدلالية في الاتجاه المعاكس بالاعتماد على الشكل النحوي ، أو يمكن أن تكون مستقلة تماما عن التبعية النحوية
Semantic Disambiguation	إزالة اللبس الدلالي	الهدف من هذه المهمة توضيح المعنى الدلالي للجمل عن طريق توضيح العلاقات بين الكلمات استنادا إلى معناها والتخلص من الغموض الذي يتخللها
Semantic Features	السمات الدلالية	هي الميزات والخصائص ذات المعنى الدلالي
Semantic Knowledge Base	قاعدة المعرفة الدلالية	هي مجموعة من الممارسات التي تسعى لتصنيف المحتوى بحيث يمكن الوصول إلى المعرفة التي يحتويها على الفور وتحويلها إلى الجمهور المطلوب ، بالشكل المطلوب. هذا التصنيف للمحتوى له خاصية في طبيعته – تحديد المحتوى حسب نوعه أو معناه ضمن المحتوى نفسه وعبر بيانات وصفية
Semantic Machine Learning System	نظام تعلم آلي دلالي	هو نظام بناء الهياكل التي تقارب المفاهيم من مجموعة كبيرة من الوثائق ،وتصنيفها عن طريق تقنيات تعليم الآلة.

Semantic Metadata	بيانات واصفة للدلالة	هي البيانات الوصفية التي تصف المحتوى أو المعلومات الخاصة بالمحتوى ذات الصلة بالمحتوى استنادًا إلى نموذج بيانات وصفية مشتركة.
Semantic Query Expansion	توسيع الاستعلام الدلالي	أحد أنواع تقنيات توسعة الاستعلام التي تعتمد على المعنى الدلالي
Semantic Web	الويب الدلالي	هو مصطلح يشير إلى الشبكة الدلالية، يطلق عليها أحياناً "الويب ذات الدلالات اللفظية" أو "الويب ذات المعنى") هو ثورة جديدة في عالم الويب حيث تصبح المعلومات والبيانات قابلة للمعالجة منطقياً من برامج الحاسوب بحيث تتحول تلك المعلومات والبيانات إلى شبكة بيانية.
Semantics	الدلالة	هي الدراسة اللغوية والفلسفية للمعنى، في اللغة ، ولغات البرمجة ، والتسجيلات الرسمية. وهي تتعلق بالعلاقة بين الإشارات–مثل الكلمات والعبارات والعلامات والرموز–وما ترمز إليه ، وهي الدلالة.
Semi-Automatic	شبه آلي	هو نظام يتم تشغيله آلياً يدوياً معًا
Semitic	دلالي	استخدم المصطلح لأول مرة في الثمانينيات من القرن الثامن عشر من قبل أعضاء مدرسة غوتنغن للتاريخ ، الذين استمدوا الاسم من شيم ، أحد أبناء نوح الثلاثة في سفر التكوين.
Semitic Language	لغة سامية	هي فرع من عائلة اللغات الأفروآسية التي نشأت في الشرق الأوسط. يتحدث لغات سامية أكثر من ٣٣٠ مليون شخص في معظم أنحاء غرب آسيا وشمال أفريقيا والقرن الأفريقي ، وكذلك في مجتمعات مغتربة كبيرة في كثير من الأحيان في أمريكا الشمالية وأوروبا.
Sentiment	حكم أو رأي	المشاعر والأحاسيس التعبيرية حول حدث معين
Sentiment Analysis	تحليل الرأي	الكشف التلقائي للآراء المنصوص عليها في النص

Sentiment Polarity	قطبية الآراء أو الميول	هي تحديد ما إذا كان الرأي المكتوب إيجابيًا أو سلبيًا أو محايدًا مع ذكر مقدار الميل من مئة فنقول ٦٠, كان إيجابيًا وهكذا
Sequence Transcription	ترجمة صوتية متسلسلة	ترجمة المقاطع الصوتية ذات التسلسل المحدد وقراءتها لتسهيل العمل عليها
Sign Language	لغة الإشارة	هي اللغات التي تستخدم الاتصال اليدوي لنقل المعنى. ويمكن أن يشمل ذلك توظيف إيماءات اليد والحركة وتوجيه الأصابع والأذرع أو الجسم وتعبيرات الوجه لتوصيل أفكار المتحدث. غالبًا ما تشترك لغات الإشارة في أوجه تشابه كبيرة مع لغتها المنطوقة.
Sign Writing	كتابة الإشارات	هو نظام كتابة لغات الإشارة. وهو نظام مميز للغاية ومبدع بصريًا ، سواء في أشكال الشخصيات ، أو الصور المجردة للأيدي والوجه والجسم ، وفي ترتيبها المكاني على الصفحة ، والتي لا تتبع أمرًا تسلسليًا مثل الحروف التي تجعل حتى الكلمات الإنجليزية مكتوبة.
Silent Letter	الحرف الصامت	هي أحرف معينة في كلمات محددة لا تتوافق مع أي صوت في نطق الكلمة. وتتطلّب الكتابة الصوتية التي تُصوِّر بشكل أفضل النّطق والملاحظة التي تتغيّر بسبب القواعد وقرب الكلمات الأخرى رمزًا لإظهار أن الحرف كتم الصوت.
Singular	المُفرد	هي مفهوم يدل على نقيض الجمع ويدل على أقل من اثنين
Singular Value Decomposition (SVD)	تفكيك القيم المنفردة في المصفوفات	هو واحد من معاملات المصفوفة الحقيقية أو المعقدة. وهي تقنية قوية للتعامل مع مجموعات من المعادلات أو المصفوفات التي تكون إما مفردة أو معدودة، وتكون قريبة جدًا من المفرد. وتستخدم في العديد من التطبيقات المفيدة في معالجة الإشارات والإحصاءات.

Skeleton	**الهيكل**	**هو الشكل الأساسي من العنصر أو أصغر وأدق تمثيل للمجسم أو المحتوى البياني**
Smoothing	**تنعيم**	**هي مهمة تنظيف الصورة وتنعيمها وإظهارها بشكل أفضل عن طريق استخدام طرق خاصة للتخلص من الضوضاء**
Software Agents	**البرنامج الوكيل**	**هو برنامج حاسوبي يعمل لمستخدم أو برنامج آخر في علاقة وكالة، (agere الشتقة من اللاتينية) والتي تعني القيام) وتعني أيضا الاتفاق للتصرف نيابة عن.**
Source-Channel Model	**نموذج القناة المصدرية**	**نموذج إحصائي لحل مشاكل الغموض في التحليل**
Source Language	**اللغة المصدر**	**في الترجمة الآلية، اللغة التي يتم الترجمة منها**
Special Marks	**علامات خاصة**	**هي مجموعة من الرموز الدلالية التي تهدف لتمييز صفة معينة**
Speech Corpora	**الذخيرة الكلامية**	**هي معاجم ومخازن تحتوي بيانات صوتية يتم تسجيلها من موارد مختلفة مثل تسجيل صوت البشر**
Speech Recognition	**التعرف إلى الكلام**	**هي طريقة التعرف واكتشاف الكلام المنطوق وتحويله إلى مكتوب**
Speech Recognition Applications	**تطبيقات التعرف إلى الكلام**	**هو المجال الفرعي متعدد التخصصات من اللغويات الحاسوبية التي تطور منهجيات وتقنيات تمكن من التعرف إلى اللغة المنطوقة وترجمتها إلى نص بواسطة أجهزة الحاسوب. وهو يشتمل على المعرفة والبحث في علم اللغة وعلوم الحاسوب ومجالات الهندسة الكهربائية.**
Speech Signal	**إشارات الكلام**	**إشارات موجية كهربائية تمثل موجة الصوت.**
Speech Synthesis	**توليد الكلام**	**إنشاء صوت يحاكي لغة البشر باستخدام الحاسوب**

Spell Checker	المدقق الإملائي	هو البرنامج الذي يدقق النّصوص كلمة بعد أخرى عن طريق مقارنتها بقاعدة بيانات محفوظة في ذاكرته لكلّ الكلمات الممكنة، أما الكلمات التي لا يعرفها فإمّا أن يقترح البديل أو يترك الخيار للمستعمل في استبدالها أو المحافظة عليها وإدخالها في قاموس المستعمل.
Spelling	تهجئة	هو مزيج من الحروف الأبجدية لتشكيل كلمة مكتوبة. إنها عملية لغوية من الكتابة الصحيحة مع الحروف والتشكيلات اللازمة الموجودة في ترتيب مفهومي عادة.
Standard Arabic Morphological Analyzer (SAMA).	المُحلّل الصّرفي للعربية القياسية	برنامج للتّحليل الصّرفيّ يقوم بتعرّف مكوّنات الكلمات مثل جذورها وجذوعها وتفكيكها إلى سوابق ولواحق وأواسسط. وتستفيد هذه البرامج كثيرا من الشّكل في عمليّة التّحليل لأنّ ذلك يساعد على تحديد أيّ قسم من من أقسام الكلام تنتمي إليه الوحدة المحلّلة.
Standardization	المعايرة	هي محاولة توحيد لمجموعة عناصر بحيث تندرج تحت مقاييس ومعايير محددة وثابتة
State	الحالة	هي حالة منظمة تصمم في ظل هيكل واحد وتتبع مجموعة ثوابت محددة وتعبر عن خط سير محدد لتطبيق هدف واضح.
Statistical Approach	المقاربة الإحصائية	هي عبارة عن استخدام الطرق الرقمية والرياضية فى معالجة وتحليل البيانات وإعطاء التفسيرات المنطقية المناسبة لها ويتم ذلك عبر عدة مراحل منها جمع البيانات الإحصائية عن الموضوع. وعرض هذه البيانات بشكل منظم وتمثيلها بالطرق الممكنة وتحليل البيانات وتفسيرها.

Statistical Machine Translation (SMT)	الترجمة الآلية الإحصائية	هو نموذج للترجمة الآلية حيث يتم إنشاء الترجمات على أساس النماذج الإحصائية التي يتم اشتقاق معاملاتها من تحليل النصوص ثنائية اللغة. ويتناقض النهج الإحصائي مع الأساليب المستندة إلى القواعد للترجمة الآلية بالإضافة إلى الترجمة الآلية المستندة إلى المثال.
Statistical Measure	مقياس إحصائي	هو مقياس لحساب تفهرس عائلة من التوزيعات الاحتمالية. ويمكن اعتباره خاصية عددية لمجموعة سكانية أو نموذج إحصائي.
Statistical Models	النماذج الإحصائية	هي فئة من النماذج الرياضية ، التي تجسد مجموعة من الافتراضات المتعلقة بتوليد بعض بيانات العينة ، وبيانات مماثلة من مجموعة أكبر من السكان. يمثل النموذج الإحصائي ، غالباً في شكل مثالي إلى حد كبير، عملية توليد البيانات.
Stem	الجذع	هو جزء أساسي من كلمة لا تشمل أي إضافات صرفية
Stem Based (word based)	على أساس الكلمة الجذع	هي التقنيات التي تعتمد على جذع الكلمة وليس جذرها وأصلها
Stemmer	مجذع	برنامج يعيد الكلمة إلى حالتها المستخدمة بإزالة السوابق واللواحق منها مما يعيدها إلى جذعها أو إلى الجذر مع بعض الزيادات.
Stemming	استخراج جذع الكلمة	عملية تحليل الكلمة لاستخلاص المحتوى الأصلي منها واستخلاص حالتها الحقيقية في الاستخدام دون زيادات.
Stemming Algorithm	خوارزمية التجذيع	خوارزميات تتضمن خطوات محددة تهدف إلى استخلاص واستخراج جذع الكلمة الحقيقي وذلك حسب قواعد اللغة او جذرها المعنوي وذلك بالتخلص من أحرف المعاني الزائدة على أصل الكلمة

Stemming Technique	تقنية إيجاد جذع الكلمة	تقنيات مطورة تسعى إلى الوصول إلى أفضل طريقة للتخلص من الزوائد الحرفية على الكلمة واكتشاف الجذع الأصلي لها
Stop Words	الكلمات المستبعدة	هي مجموعة من الكلمات والحروف التي يتم ترشيحها قبل أو بعد معالجة بيانات اللغة الطبيعية.
String	سلسلة رمزية	سلسلة مكونة من رموز أو حروف وعادة ما تُعدُّ الكلمات عى أنها سلاسل رمزية
String Matching Approach	طريقة مطابقة الحروف	هو إجراء لفحص تسلسل محدد من الرموز والأحرف والأرقام لوجود مكونات بعض النماذج، ويجب أن تكون المطابقة دقيقة. ونماذج مقارنة النصوص عموما لها شكل إما تسلسلي أو شجري
Structural Ambiguity	اللبس الدلالي للجمل	هي الحالة التي يمكن فيها تفسير الجملة بأكثر من طريقة واحدة بسبب بنية الجملة الغامضة. لا ينشأ الغموض التراكيبي من نطاق المعاني للكلمات المفردة ، ولكن من العلاقة بين الكلمات والعبارات في الجملة.
Subjunctive	شرطي	عادةً ما تُستخدم أشكال الأفعال الشرطية للتعبير عن حالات عدم واقعية مختلفة مثل الرغبة أو الانفعال أو الاحتمال أو الحكم أو الرأي أو الالتزام أو الفعل.
Subtree (ST)	الفروع	مصطلح يدل على فئة فرعية تنحدر من سلالة عائلية محددة ويعبر عنها بشكل هيكل شجري
Sub-Word Segmentation	تجزئة الكلمة إلى أجزائها	هي عملية تقسيم الكلمة إلى أجزاء أصغر لاستخراج جذورها الحقيقية وتقليمها وتحديد زوائدها الحرفية
Suffix	لاحقة	اللواحق التي تأتي في آخر الكلمة
Summarization	تلخيص	هو عبارة عن تقنية إعادة وتمثيل محتوى معين بأقل كم ممكن من الكلمات مع الحفاظ على جوهر المحتوى الأصلي

Summary	مُلخص	هو ناتج عملية التلخيص وهو جزء معين يعبر عن محتوى معين بكم قليل من الكلمات بحيث تتضح الفكرة من المحتوى الأصلي
Summary Evaluation	تقييم التلخيص	مرحلة تقييم أداء دقة عملية التلخيص ومقدار دقة تمثيلها للمحتوى الأصلي
SUMO Ontology	الأنطولوجيا العُليا المُدمَجة المُقترحة	الأنطولوجيا العلوية يقصد بها أنطولوجية الأساس لمجموعة متنوعة من أنظمة معالجة معلومات الحاسوب. تحدد SUMO التسلسل الهرمي للفصول والقواعد والعلاقات ذات الصلة. يتم التعبير عنها في نسخة من لغة SUO-KIF التي لها بناء يشبه LISP.
Sun Letter	الحروف الشمسية	في اللغة العربية، تنقسم الحروف إلى مجموعتين، الحروف الشمسية والحروف القمرية استنادًا إلى ما إذا نطق الحرف اللام من (أل) التعريف بعد دخولها على الاسم أم لا، في الحروف الشمسية لا تلفظ اللام، ويشدد الحرف بعد (أل) التعريف.
Superlatives	صيغ التفضيل	صيغة تتعلق أو تشكل درجة المقارنة النحوية التي تشير إلى مستوى أو حد أقصى متطرف أو غير مسبوق، أو التي تشير إلى من تجاوز جميع الآخرين وامتلك الأفضلية عليهم في سلوك معين سواء حسن أو قبيح
Supervised Learning	التعلم آلي مراقب	نموذج للتعلم الآلي يعتمد على بيانات للتدريب بحيث يبقى تحت الإشراف.

Support Vector Machine (SVM)	آلية تحليل للبيانات	هي إحدى خوارزمات تعلم الآلة المراقبة (تحت الإشراف) لتحليل البيانات من أجل تصنيفها إحصائيا وعمل تحليل الانحدار اللازم لها. بحيث تقوم هذه الخوارزمية بإيجاد إطار خطي للفصل بين خصائص كلٍ من البيانات المدخلة بحيث تكون الهوة بينهما متسعة قدر الإمكان.
Surface Patterns	الأوزان	قائمة بجميع الأوزان الأساسية مثل فعل، فعول، مفاعيل،... إلخ
Survey	مسْح أو دراسة مسحية	هي دراسة مسحية تهدف إلى تعمق الاطلاع والاستطلاع حول موضوع معين مع الأخذ بعين الاعتبار أغلب التجارب التي جرت عليه
Syllable	مقطع صوتي	مقطع صوتي ينطق منفصلاً، قد يتكون من حرف واحد متحرك، أو حرفين؛ الأول متحرك والثاني ساكن، فكلمة (مِنْ) تتكون من مقطع صوتي واحد، وكلمة (كَتَبَ) تحتوي على ثلاثة مقاطع صوتية.
Syllable-Based Morphology	الصرف المعتمد على الوحدات اللفظية للكلمة	التحيل الصرفي للكلمات الذي يعتمد الوحدات اللفظية (Syllables) كأساس للتحليل.
Symbolic Approach	المقاربة الرمزية	يستخدم المنهج الرمزي لتمثيل المعرفة ومعالجتها الأسماء لتعريف معنى المعرفة الممثلة بوضوح. يتم وصف المعرفة الممثلة من خلال الأسماء المعطاة للجداول والحقول والطبقات والسمات والطرق والعلاقات إلخ.
Synchronized Computational Model	النموذج الحاسوبي المتزامن	هو النموذج المحوسب لحل مشاكل معينة أو إنجاز مهمات محددة بشكل متزامن مع بعضها

Synonyms	المترادفات	المترادفات كلمتان مختلفتان أو أكثر يمكن استخدامها في تمثيل نفس المفهوم أو المعنى في اللغة. هناك نوعان من المترادفات المترادفات المطلقةsynonyms absolute))وهي التي لها معان متطابقة من جميع النواحي وفي جميع السياقات، المترادفات الجزئية التي تتمايز جزئيًا مع إمكانية استخدامها في السياقات ذاتها.
Synsets	مجموعات المترادفات	مجموعة من المترادفات القابلة للتبادل في بعض السياقات دون تغيير قيمة الحقيقة التي تتضمنها.
Syntactic	نحوي	هو مفهوم يتعلق بالبناء اللغوي للكلمات في الجملة أو العبارة، يهدف إلى ضبط قواعد التركيب لإقامة الدلالة في الجمل بطريقة لا تحتمل اللبس أو الخطأ.
Syntactic Ambiguity	الغموض النحوي	هو وجود أكثر من تحليل محتمل لجملة معينة، بحيث تتعدد وظيفة الكلمة فيها بما ينتج عنه تعدد في الدلالة قد يصل حدّ التناقض في المعنى.
Syntactic Analysis	التحليل النحوي	هي عملية تحليل سلسلة من الرموز ، سواء في اللغة الطبيعية أو لغات الحاسوب أو هياكل البيانات ، بما يتفق مع القواعد الرسمية. في نظام تركيزي محدد.
Syntactic Annotation	الوسم النحوي	هو شكل من أشكال البيانات الوصفية النحوية التي يمكن إضافتها إلى المعرفة، وقد يتم وضعها على شكل علامات على الفئات والطرق والمتغيرات والمعلمات والحزم.
Syntactic Dependency Structure	هيكلية الاعتماد النحوية	يتم تمثيل أي تبعية نحوية بين كلمتين بشكل عام في الفضاء الدلالي كإحالة بناء على علاقة الكلمة بالفعل الذي يمثل مركز الجملة.
Syntactic Form	الشّكل النحوي	نحويا كل كلمة تقسم إلى فئات وأشكال مختلفة حسب قواعد وأسس معينة

Syntactic Parses	المحلل النحوي	هو البرنامج الذي يقوم بتحليل الجملة (أو إعرابها) آليا، فبعكس تصنيف أقسام الكلام الذي يتم على مستوى الكلمة الواحدة في التحليل الصرفي، يتم في التحليل النحوي النظر إلى الجملة ككل ومن ثم تحليلها (أو إعرابها)
Syntactic Properties	الخصائص النحوية	هي عبارة عن مجموعة من الميزات والسمات النحوية التي تميز تركيب الجملة والكلمات التي تشكلها
Syntactic Structure	تركيب نحوي	علاقة تجمع لفظتين فأكثر على أساس وظيفي دلالي، يظهر أثر هذه العلاقة بعلامات صوتية وقيم دلالية تشكل الحد الأدنى للجمل والنصوص.
Syntax	النحو	علم النحو، دراسة التوافقية لمفردات اللغة في حالة التركيب (دون الإشاره إلى معناها).
Syntax Analyzer	المُحلل النحوي	هو عبارة عن نموذج يستخدم لتقسيم الجملة إلى كلمات كل كلمة حسب نوعها ووظيفتها النحوية
Syntax Tree	التشجير النحوي	هو عبارة عن مقطع هيكلي شجري يعبر عن العلاقة النحوية التي تربط الكلمات التي تكون الجملة
Synthetic Speech Waveforms	موجات الكلام الاصطناعية	واحد من أهم التقنيات الأساسية التي تولد أشكال موجة تركيبية اصطناعية، والاستخدامات المقصودة لنظام التوليف سوف تحدد عادة النهج الذي يتم استخدامه
Tag	وسم	هي رمز يضاف إلى الكلمة في النص ويستخدم لتصنيف نوعها الصرفي أو النحوي
Tag Set	مجموعة وسم	قائمة الوسوم المستخدمة في تعليم كلمات نص.
Tagging	الوسم	عملية وسم النصوص.
Target Language	اللغة المستهدفة	في الترجمة الآلية، اللغة التي يتم الترجمة إليها
Taxonomy	تصنيف إلى شجرة الموضوعات	هو تصنيف أو تقسيم مجموعة من العناصر إلى فئات حسب التشابه والاختلاف فيما بينها

Tense	زمن	الزمن للفعل، ماضٍ، أو حاضر أو مستقبل
Term Extraction Approaches	طرق استخراج المصطلحات	تعالج هذه البرامج النّصوص للتّعرّف إلى المصطلحات انطلاقا من خاصيّاتها التّركيبيّة في الجملة أو من مؤشّرات أخرى كورودها بين ظفرين أو بالبنط الغامق أو بحرف بارز. وعادة ما يتطلّب العمل الآليّ تدخّل المستعمل لتحديد الألفاظ المرشّحة للمعالجة أو المراجعة.
Term Frequency – Inverse Document Frequency (TF-IDF)	نسبة تردد المصطلح إلى عدم ظهوره في المستند	هي اقتران يعبر عن نسبة ظهور مصطلح معين في أحد المستندات إلى عدم ظهوره في بقية المستندات
Text Clustering	عنقدة النص	هي عملية تجزئة النص وتقسيمه إلى مجموعات بحيث تتشارك كل مجموعة بصفات محددة
Text Encoding Initiative (TEI)	مبادرة ترميز النص	منظمة غير ربحية تتكون من مؤسسات أكاديمية ومشاريع أبحاث وعلماء من مختلف أنحاء العالم، تطور بشكل جماعي معيارًا لتمثيل النصوص في شكل رقمي. إن هدفها الرئيسي هو مجموعة من الإرشادات التي تحدد طرق التشفير للنصوص المقروءة آليًا ، وبشكل أساسي في العلوم الإنسانية والعلوم الاجتماعية واللغويات. منذ عام ١٩٩٤ ، تم استخدام إرشادات TEI على نطاق واسع من المكتبات والمتاحف والناشرين والباحثين الأفراد لتقديم نصوص للبحث والتعليم والحفظ عبر الإنترنت.
Text Mining	التنقيب في النصوص	هي مهمة التنقيب في النصوص لاستخراج معلومات ومعارف معينة منها للمساعدة في اتخاذ القرارات
Text Normalization	تطبيع النص	هي مهمة تهدف لإعادة تمثيل النصوص بشكل يبسط التعامل معها

Text Processing Component	مكون معالجة النصوص	هي الخطوات الأساسية التي تتألف منها مرحلة المعالجة الأولية للنصوص والتي تتضمن التخلص من كلمات الوقف وتقليم الكلمات وتجذيرها وتوحيد وتتطبيع كتابة بعض الأحرف
Text Recognition	تمييز النصّ	هو النظام الذي يستخدم للتعرف إلى النصوص المكتوبة باليد أو النصوص المصورة وتحويلها إلى مطبوعة قابلة للتعديل
Text Segmentation	تجزئة النص	هي مهمة تهدف إلى تقسيم وتجزئة النصوص إلى كلمات
Text Similarity Techniques	تقنيات تشابه النص	هي الوسائل أو الطرق التي تستخدم لحساب مقدار التشابه بين النصوص بناء على محتواها من الكلمات
Text Summarization	تلخيص النصوص	هي عملية إعادة تمثيل محتوى نصي معين بكلمات أقل مع الحفاظ على المعنى
Text-Based Model	نموذج اعتماد النص	نموذج قائم على النصوص في مقابل نموذج قائم على الصور
Text-Speech Synthesizer	مولد الكلام من النص	محول النص إلى كلام (TTS) يُعدُّ نوعًا من تطبيقات تركيب الكلام التي يتم استخدامها لإنشاء إصدار صوت منطوق للنص في مستند حاسوبي، مثل ملف مساعدة أو صفحة ويب. يستطيع TTS قراءة معلومات معروضة على الحاسوب للشخص الذي يواجه تحديًا بصريًا، أو غير ذلك بشكل آلي.
The Ambiguity Pyramid Hypothesis	فرضية الغموض الهرمية	تفترض فرضية الغموض الهرمية أن النظام الغني والمعقد يؤثر إيجابا على الأداء حيث يساعد على تقليل الغموض بدلاً من زيادته.

The Arabic Parser	المعرب اللغوي العربي	التّعرّف الآليّ إلى أقسام الكلام بالنّسبة إلى كلّ وحدة معجميّة في النّصّ. لا يتعلّق الأمر بأقسام الكلام الثّلاثيّة، كما يعرفها العرب (اسم وفعل وحرف) بل يتجاوزها التّحليل إلى تحديد العدد (جمع أو مثنّى أو مفرد) والجنس (مؤنّث أو مذكّر أو ملتبس) وكذلك زمن حدوث الفعل...
The Semantic Web	الشبكة الدلالية	يطلق عليها أحياناً "الويب ذات الدلالات اللفظية" أو "الويب ذات المعنى") هو ثورة جديدة في عالم الويب حيث تصبح المعلومات والبيانات قابلة للمعالجة منطقياً من برامج الحاسوب بحيث تتحول تلك المعلومات والبيانات إلى شبكة بيانات ذات معنى.
Thesauruses	المكانز	المكنز هو عمل مرجعي يسرد الكلمات مجمعة معًا وفقًا لتشابه المعنى (يحتوي على مرادفات وأحيانًا متضادات) ، على النقيض من القاموس ، الذي يوفر تعريفات للكلمات ، ويدرجها عمومًا بترتيب أبجدي. الغرض الرئيسي من مثل هذه الأعمال المرجعية للمستخدمين هو العثور على الكلمة.
Thinning	الترقيق والتنحيف	هي إزالة الزوائد والتخلص من الإضافات للحصول على التمثيل الأكثر ترابطا
Time and Place Adverbs	ظروف المكان والزمان	هي الأسماء التي تحدد زمن أو مكان الخبر المنقول في الجملة.
Token	مدخل معجمي	حدث فردي لوحدة لغوية في الكلام أو الكتابة.
Tokenization	التأخيذ	هي تقنية تقسيم النص إلى كلمات اعتمادا على الفواصل البيضاء أو الفراغات بين الكلمات
Tokenizer	محلل معجمي	هو النموذج الذي يقوم بمهمة تقسيم النص وتجزئته
Training Corpus	ذخيرة التدريب	ذخيرة لغوية تستعمل لتدريب خورزميات التعلم الآلي
Transducer	محول	برنامج تحويل

Transducer Cascade	المحول التتابعي	هو ما يحول الطاقة من صورة ميكانيكية إلى كهربية بشكل تتابعي
Transition	انتقالي	هو محول الطاقة(محول الإشارة) الذي يستخدم في حالات كثيرة وغالبا عندما تكون الكمية المراد قياسها غير كهربية لتسجيل أو تداول الكمية المقاسة في عمليات التحكم حيث يلزم غالبا تحويلها إلى كمية كهربية
Transliterated	مترجم صوتياً	الترجمة الصوتية هي نوع من تحويل النص من نص إلى آخر وهي كلمة منحوتة من كلمتين، نقل وحرف، وتعني نسخ الحروف ورسمها بنظام كتابة آخر، أي إيقاع تقابل بين لغتين ومبادلة كل حرف بحرف واحد كلما أمكن. فهو محاولة للتوسط بين المنطوق بلغة والمكتو بلغة أخرى.
Tree Edit Distance (TED)	مقياس التشابه التشجيري	هو تسلسل أدنى تكلفة لعمليات العقدة التي تحول نظام تشجيري إلى آخر، ويعبر عن مقياس مسافة معروفة للبيانات الهرمية. ويستخدم في مجموعة واسعة من التطبيقات المتنوعة مثل هندسة البرمجيات ، ومعالجة اللغات الطبيعية ، والمعلوماتية الحيوية.
Tree Structure	الهيكل تشجيري	هو هيكل تنظيمي تتابعي يشبه تركيبكة الشجرة بحيث يبدأ من الآباء الى الأبناء ومن الأصول إلى الفروع
Treebank	بنك التشجير النحوي	هو عبارة عن ذخيرة موسومة بعلامات تبين التركيب النحوي لجملها للاستفادة منها في عرض الذخيرة على شكل أشجار تمثل العلاقات النحوية، مجموعة لغوية تم إعراب جملها بحيث إن البناء النحوى للجملة يمكن التعبير عنه فى شكل شجرة
Tri-Consonant	ثلاثي الحروف	تركيبة لغوية تتكون من ثلاثة حروف تشكل من كلمة أصلية أو تراكيب لغوية مختلفة

Trigger Words	الكلمات المؤثرة	هي كلمات دلالية مرمزة ومميزة وتكون معلمة للإرشاد والدلالة على حدث معين
Ubiquitous	واسع الانتشار	يظهر في كل مكان وفي وقت واحد ؛ او هو ما يحدث أو يبدو في أكثر من موقع واحد في نفس الوقت.
Unicode	نظام ترميز الحروف الدولي الموحد	مشروع معيار الشيفرة الموحدة الذي يهدف إلى تطوير ترميز يدعم جميع أحرف لغات العالم.
Unified Modeling Language (UML)	لغة النّمذجة الموحّدة	هي لغة قياسية لتحديد ، وتصوير ، وبناء ، وتوثيق الأعمال لبرمجيات الأنظمة ، فضلا عن نماذج الأعمال التجارية وغيرها من النظم المختلفة عن البرمجيات.
Uniform Resource Locator (URL)	رابط الموقع على الشبكة	هو اختصار للكلمة الإنجليزيّةUniform Resource Locator،وتعني عنوان الإنترنت؛ فالشريط الموجود على المتصفّح للذهاب أو الدخول على موقع معيّن يضم http://، فعلى سبيل المثال عنوان موقع موضوع http://www.example.com يضم التالي: البروتوكول: وهو بروتوكول الإنترنت...
Universal Networking Language (UNL)	لغة التشبيك العالمية	هي لغة رسمية إعلامية مصممة خصيصا لتمثيل البيانات الدلالية المستخرجة من نصوص اللغة الطبيعية. ويمكن استخدامه كلغة محورية في أنظمة الترجمة الآلية باللغات أو كلغة تمثيل المعرفة في تطبيقات استرجاع المعلومات.
Unmarked Stem	جذع مجرد	جذع الكلمة بدون تغييرات تصريفية أو زوائد إلصاقية
Unstructured Data	البيانات غير المهيكلة	هي البيانات التي لا تتبع تنظيمًا محددًا أو هيكلة ثابتة

Unsupervised Approach	مقاربة غير مدارة	التعلم الآلي غير الخاضع للرقابة هو مهمة التعلم الآلي لاستنتاج وظيفة لوصف بنية خفية من البيانات "غير المسماة".
Unsupervised Learning Paradigm	نموذج التعلم غير الخاضع للرقابة	هو النموذج الذي لا يتطلب التدريب على عينة معينة قبل التنفيذ بل ينفذ المهمة عن طريق معادلات حسابية لحساب الفروقات بين جميع لعينات المتوفرة.
Unvoweled	غير مشكولة	لايوجد تشكيل على أي حرف في الكلمة
Utterances	الكلام أو الحديث	في تحليل اللغة المحكية ، الكلام هو أصغر وحدة من الكلام. هو عبارة عن جزء من الكلام المستمر يبدأ وينتهي بوقفة واضحة. في حالة اللغات الشفوية ، عادةً ما تكون مقيدة دومًا بالصمت.
Validation	التحقق	هو تقييم وتحكيم أداء نظام معين عن طريق فحصه على مجموعة معينة من البيانات وعرض مقدار دقته وسرعته وغيرها من المقاييس
Vector Feedback	التغذية الراجعة الاتجاهية	هو عملية تحويل الأنظمة غير الخطية إلى أنظمة خطية لتسهيل التعامل معها والإفادة منها
Vector Space Model (VSM)	نموذج تمثيل الوثائق على صورة متجه	هو نموذج جبري لتمثيل الوثائق النصية (وأي كائنات بشكل عام) كناقلات للمعرّفات، على سبيل المثال، مصطلحات الفهرس. يتم استخدامه في تصفية المعلومات، واسترجاع المعلومات والفهرسة والتصنيفات الملائمة. وكان أول استخدام لها في نظام SMARTلاسترجاع المعلومات.

Verb Phrase	العبارة الفعلية	هي الجملة التي تحتوي على مجموعة من الكلمات التي تشتمل على الفعل وتكملته، وعناصر الجملة الأخرى، أو غيرها من الإضافات التي تعمل بشكل تركيبي كفعل. وفي اللغة الإنجليزية الجملة الفعلية تجتمع مع اسم أو مع الجملة الاسمية لتشكيل جملة بسيطة.
Vocabulary Continuous Speech Systems	أنظمة تمييز الكلام المتواصل	هو النظام المعني باستخراج المفردات والتعرف إليها من الكلام المحكي بشكل مستمر ٍ بحيث تكون مخرجات النظام جميع المفردات التي وردت في الحديث المحكي
Vocal	صوتي	هو مصطلح يعبر عن المحتوى الصوتي ويستخدم مجالات مختلفة تتعلق بالرنين، بما في ذلك: التضخيم، والترشيح، والإثراء، والتوسع، والتحسين، والتكثيف، والإطالة.
Vocalization	1: نطق ٢: تشكيل	أولاً: طريقة إصدار صوت ما للتعبير عن كلمة معينة ثانياً: استعمل مجازاً للتعبير عن التشكيل في اللغة العربية
Voice	المبني للمعلوم أو المجهول	يكون الفعل ذا فاعل معروف في السياق ويكون عندها الفعل مبنياً للمعلوم، أو فاعله غير معروف ويكون مبنياً للمجهول
Voting Methods	طرق التصويت	هي الأنظمة التي تعتمد على مبدأ التصويت لاتخاذ قرار معين ويشيع استخدامها في أنظمة التعرف والتصنيف حيث يسمح لأكثر من مصنف باتخاذ القرار حول التعرف على نمط معين
Vowelized	مشكول	نص مشكول تظهر على كلماته الحركات.
Vowelized Document	وثيقة مشكولة	هو المستند الذي يبرز فيه مكان ظهور الحركات.

Vowels	حُروف العلة	هو الحروف التي تمثل صوت الكلام الناتج عن الأوتار الصوتية المفتوحة، وتشمل في العربية الحروف (ا، و، ي)
Warping	الانعطاف	ميزة تدعمها العديد من المعالجات ،التي تمكننا من تحديد محتوى معين باحاطة تمييزية
Waveform	شكل الموجة	الهيئة التي تظهر بها الموجة، هو شكل إشارة مثل موجة تتحرك في المتوسط المادي أو تمثيل تجريدي.
Web Ontology Language (WOL)	لغة أنطولوجيا الشبكة	هي لغة الويب الدلالي التي تم تصميمها لمعالجة ودمج المعلومات عبر شبكة الإنترنت، بحيث تم صياغة المعنى منها بطريقة مشابهة للعقل البشري. والغرض منها هو تسهيل التفسير بين المحتوى على شبكة الإنترنت من المفردات والتشكيلات التي تسمح بمعالجة الآلة التلقائية
Web Ontology Language Description Logics (OWL-DL)	المنطق الوصفي للغة أنطولوجيا الويب	هي لغة الويب الدلالي التي تعتمد على الوصف المنطقي للمحتوى اللغوي
Web-Based Corpora	الذخيرة المنشورة على الشبكة	هي المعاجم التي تعتمد على محتويات شبكة الويب والمواقع الإلكترونية للتعرف إلى مفاهيم المفردات والعلاقات بينها
Website/Websites	موقع / مواقع الشبكة	هو مجموعة صفحات ويب مرتبطة ببعضها ومخزنة على نفس الخادم. يمكن زيارة مواقع الويب عبر الإنترنت بفضل خدمة الويب ومن خلال برنامج حاسوبي يدعى متصفح الويب
Weighted Edit Distance	درجة التشابة المقاسة بين البيانات	هي العملية التي تقوم بحساب مسافة مبدئية بين سلسلتين مختلفتين وبعد ذلك يتم تعديل الأوزان حسب التكلفة أو معيار معين

Weighted Genetic Algorithms	الخُوارزمية الجينية الموزونة	هي إحدى أنواع الخوارزمية الجينية التي تعتمد على استخدام الأوزان لبناء الوظيفة الموضوعية(objective function).ويتم تغيير الأوزان بشكل متكرر خلال وقت التشغيل للسماح بتغيير اتجاهات البحث لاكتساح مساحة الحل بالكامل. ولأداء هذه المهمة، يتم إنشاء أرقام عشوائية
Windowing	الإنفاذ	يعرف نظام النوافذ على أنه النظام الذي يدير أجزاء مختلفة بشكل منفصل من شاشات العرض بحيث يتم تقاسم موارد عرض الواجهات الرسومية للبرامج بين تطبيقات متعددة في نفس الوقت (فتح أكثر من واجهة برمجية معا والتنقل بينها.(
Windowing & Framing	التأطير والإنفاذ	ربط مفهوم التأطير بالعمليات التنظيمية التي تجرى على مجموعة من البيانات بهدف الإشارة إلى أحداث دلالية معينة مثل تأطير حزم البيانات المنقولة عبر الإنترنت، حيث يكون التأطير بوضع عنوان المرسل والمستقبل وزمن الإرسال.
Word Affixes	زوائد الكلمة	هي مجموعة من الزوائد الحرفية والإضافات الاشتقاقية التي تضاف على الجذر الحقيقي للكلمة لتشكيل المزيد من الأشكال منها ولها ثلاثة أنواع إما سوابق وإما لواحق أو إضافات وسطية حسب مكان الإضافة
Word Alignment Model	نموذج محاذاة الكلمات	نموذج يستعمل في الترجمة الآلية لمحاذاة (مطابقة) الكلمات في جملة بلغة معينة (اللغة المصدر)، وترجمة تلك الجملة بلغة أخرى (اللغة المستهدفة)
Word Choice	اختيار الكلمات	هي مفردات محددة يستخدمها الكاتب لتوصيل المعنى وتنوير القارئ. ويتم ذلك عن طريق استخدام لغة غنية وملونة ودقيقة لا تتصل بطريقة وظيفية فقط ، ولكن بطريقة متفاعلة.

Word Classifier	مصنف الكلمات	تصنيف الكلمات إلى مجموعات حسب خصائص محددة، وتستخدم خوارميات تصنيف شهيرة لهذا الغرض.
Word Collocation Data	بيانات رصف الكلمة	هي واحدة من موارد المعلومات التي يتم التحقق فيها من اللغويات الحسابية وتُعدُّ كأداة مفيدة لعملية التعرف إلى الكلمات بحيث تعبر عن احتمال أن كلمتين تتشاركان في الحدوث ضمن نفس المجال مثل كلمة قارب ونهر.
Word Matching	مواءمة الكلمات	هي عملية مطابقة ومقارنة بين الكلمات التي تتم عن طريق حساب نسبة التشابه أو الاختلاف فيما بينها
Word Sense Disambiguation	فك اللبس الدلالي للكلمة	استخلاص المعنى الصحيح للكلمة وفقا للسياق.
WordNet	شبكة الكلمات	هي قاعدة بيانات معجمية للغة، يتم فيها تجميع الأسماء والأفعال والصفات والظروف في مجموعات من المرادفات المعرفية التي تسمىsynsets ، وتقدم تعريفات قصيرة وأمثلة لاستخدام المفردات، ويسجل عدد من العلاقات بين هذه المجموعات المرادفات أو عناصرها.
XML-based Morphological Definition Language (XMODEL)	لغة التعريف الصرفي المعتمدة على لغة التوصيف الموسعة	لغة لبناء المعاجم تعتمد على لغة التعريف الصرفي للمدخلات التي تستعمل لغة التوصيف الموسعة (XML)

الملحق 2
الكشاف الموضوعي
مسرد الدراسات الشمولية والعامة في حوسبة اللغة العربية

الرقم	عنوان الدراسة	المؤلف	مكان النشر	تاريخ النشر
١	الثورة التكنولوجية واللغة	محمد صالح عمر	دار الشؤون الثقافة العامة –وزارة الثقافة والإعلام، بغداد– العراق	١٩٨٦
	استخدام اللغة العربية في علوم الحاسوب	أحمد الأخضر غزال	المجلة العربية للتربية، تونس، المجلد٦، العدد١	١٩٨٦م
٢	اللغة العربية والحاسوب	نبيل علي	مجلة عالم الفكر، م ١٨، ع ٣	١٩٨٧
	التقنيات الحديثة واللغة العربية	محمد ظافر الصواف	مجمع اللغة العربية الأردني	١٩٨٧م
٣	اللغة العربية والحاسوب	نبيل علي	مؤسسة تعريب، الكويت	١٩٨٨
٤	اللغة العربية والحاسوب (تعليق على بحث للدكتور نبيل علي)	قاسم السارة	مجلة عالم الفكر، م١٩، ع٢	١٩٨٨
٥	آخر التطورات في مجال تقييس تعريب الحاسوب	أحمد أبو الهيجاء	مؤتمر الكويت الأول للحاسوب –جمعية الحاسب الآلي الكويتية	١٩٨٩
٦	معالجة اللغة العربية بالحاسوب	محمد عبد المنعم حشيش	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	١٩٩٢

٧	استخدام اللغة العربية في تقنية المعلومات (تسجيل مرئي)	إعداد عبد الرحمن الحاج صالح؛ تقديم محمد بن عبد الرحمن الربيع	إنتاج مكتبة الملك عبد العزيز العامة. بيانات النشر الرياض : مكتبة الملك عبد العزيز العامة	١٩٩٢
٨	بحوث ودراسات في اللسانيات الحاسوبية	عبد الرحمن الحاج صالح	موفم للنشر – الجزائر	د. ت
٩	اللسانيات وبرمجة اللغة العربية	محمد الزركان	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات– مكتبة الملك عبد العزيز – الرياض	١٩٩٣
١٠	استخدام اللغة العربية وتقنية المعلومات	إعداد محمد الحناش	مجلة التواصل اللساني. المجلد الأول	١٩٩٣
١١	العرب وعصر المعلومات – الفصل التاسع	نبيل علي	سلسلة عالم المعرفة – الكويت	١٩٩٤
١٢	الحاسوب واللغة العربية	عبد ذياب العجيلي	جامعة اليرموك الأردن	١٩٩٦
١٣	اللغة العربية والتقنيات المعلوماتية المتقدمة	محمد الحناش	وقائع المؤتمر الدولي الثاني، مجلة التواصل اللساني. المجلد الثالث	١٩٩٦
١٤	استخدام اللغة العربية في المعلوماتية	مجموعة من المؤلفين	المنظمة العربية للتربية والثقافة والعلوم تونس	١٩٩٦
١٥	الحوار إنسان – آلة	نورالدين اللوز	مجلة التواصل اللساني – مؤسسة العرفان للاستشارات التربوية والتطوير المهني – المغرب	١٩٩٦
١٦	التقنيات الحديثة وآفاقية اللغة العربية	إسحق علي حبيبي	الموسم الثقافي الرابع عشر لمجمع اللغة العربية الأردني	١٩٩٦

١٧	المشروع اللساني الحاسوبي العربي	محمد الدهبي	مجلة التواصل اللساني – مؤسسة العرفان للإستشاراتالتربوية والتطوير المهني – المغرب	١٩٩٦
١٨	اللسانيات الحاسوبية مفهومها وتطورها ومجالات تطبيقها	بلقاسم اليوبي	مجلة مكناسة ع١٢	١٩٩٩
١٩	العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية	نهاد الموسى	المؤسسة العربية للدراسات والنشر، بيروت	٢٠٠٠
٢٠	تهيئة اللغة العربية لمواجهة طوفان المعلومات والعولمة	سلوى حمادة السيد	مؤتمر لسان العرب بمقر جامعة الدول العربيةٱ القاهرة	٢٠٠٠
٢١	اللغة العربية والحاسوب	هادي نهر	مجلة التواصل (جامعة عدن) العدد ٤،	٢٠٠٠
٢٢	حوسبة التراث العربي	علي حلمي موسى	محاضرة ألقيت في مجمع اللغة العربي الأردني في ١٧ نيسان (أبريل)	٢٠٠١
٢٣	اللغة العربية والحاسوب (قراءة سريعة في الهندسة اللسانية العربية) أو مقاربة في محاكاة الدماغ العربي لغويا	محمد الحناش	محاضرة بكلية العلوم الإنسانية والاجتماعية ، قسم اللغة العربية وآدابها جامعة الإمارات العربية المتحدة. مؤسسة العرفان للاستشارات التربوية والتطوير المهني	٢٠٠٢

٢٤	التجارب الراهنة حول حوسبة النصوص التي تعتمد اللغة العربية	موسى زمولي	مجلة اللغة العربية –العدد السابع الجزائر	٢٠٠٢
٢٥	تحليل اللغة العربية بواسطة الحاسب الآلي	سعد بن هادي القحطاني	جلة علوم اللغة ، م ٥، ع ٣، (القاهرة : دار غريب.)	٢٠٠٢
٢٦	قضايا اللغة العربية في عصر الحوسبة والعولمة	نبيل علي	مجمع اللغة العربية الأردني	٢٠٠٢
٢٧	الصياغة المنطقية الخليلية وفق القواعد التوحيدية	عشيت عبد المجيد	مجمع اللغة العربية الأردني	٢٠٠٢
٢٨	دور النظرية الخليلية الحديثة في النهوض بالبحوث الحاسوبية الخاصة باللغة العربية	عبد الرحمن الحاج صالح	مجمع اللغة العربية الأردني	٢٠٠٢
٢٩	اللغة العربية والحاسوب قراءة سريعة في الهندسة اللسانية العربية	محمد الحناش	مجلة التواصل اللساني مج٩	٢٠٠٣
٣٠	اللسانيات والحاسوب واللغة العربية	مازن الوعر	صحيفة رؤى ثقافية (سورية) العدد ٤	٢٠٠٣
٣١	تمثيل الكفاية اللغوية للحاسوب	نهاد الموسى	ورقة مقدمة إلى ندوة الهوية اللغوية والعولمة جامعة البترا الأردنية الخاصة	٢٠٠٣
٣٢	الثنائيات في قضايا اللغة العربية المعاصرة الفصل من الوصف إلى التوصيف.. مقاربة في حوسبة العربية	نهاد الموسى		٢٠٠٣

٣٣	اللسانيات: المجال والوظيفة والمنهج الفصل السابع	سمير استيتية	عالم الكتب الحديث، الأردن	٢٠٠٥
٣٤	اللسانيات الحاسوبية العربية...المفهوم، التطبيقات، الجدوى	وليد العناتي	مجلة الزرقاء للبحوث والدراسات، عمادة البحث العلمي بجامعة الزرقاء الأهلية الأردنية، المجلد السابع، العدد الثاني	٢٠٠٥
٣٥	اللسانيات الحاسوبية واللغة العربية	عز الدين غازي	الحوار المتمدن ع١٦٣٩	٢٠٠٦
٣٦	الفجوة الرقمية في اللغة العربية	عبد المجيد نصير	مجمع اللغة العربية الأردني موسمه الثقافي الرابع والعشرين	٢٠٠٦
	اللسانيات الحاسوبية العربية رؤية ثقافية	وليد العناتي	مجلة فكر ونقد – المغرب	٢٠٠٦
٣٧	الحاسوب وميكنة اللغة العربية	رأفت الكمار	القاهرة : دار الكتب العلمية للنشر والتوزيع	٢٠٠٧
٣٨	دليل الباحث إلى اللسانيات الحاسوبية العربية	وليد العناتي وخالد الجبر	دار جرير للنشر والتوزيع عمان، الأردن	٢٠٠٧
٣٩	توظيف اللسانيات الحاسوبية في خدمة الدراسات العربية : جهود ونتائج	عبد الرحمن بن حسن العارف	مجلة مجمع اللغة الأردني، ع.٧٣	٢٠٠٧
٤٠	اللغة العربية والحاسوب	ندى غنيم وأميمة الدكاك	الاحتماع الثاني لخبراء المعجم الحاسوبي التفاعلي– الرياض	٢٠٠٨
٤١	اللُّغة العربيّة والشّبكة العنكبوتيّة / قضايا وحلول	أحمد حياصات	مجمع اللغة الأردني	٢٠٠٨

٤٢	المعالجة الآلية للغة العربية (المشاكل والحلول)	سلوى حماده	دار غريب للطباعة والنشر والتوزيع	٢٠٠٩
٤٣	أساسيات الحاسب المعالجة الآلية للغة العربية	شريف عصام خطاب		٢٠٠٩
٤٤	فعاليّة اللسانيات الحاسوبية العربية	ديدوح عمر	الأثر-مجلة الآداب واللغات، جامعة قاصدي مرباح، الجزائر، ع٨	٢٠٠٩
٤٥	اللسانيات الحاسوبية العربية: الإطار والمنهج	وجدان محمد صالح كنالي	المؤتمر الدولي الثاني للغة العربية - دبي	٢٠١٣
٤٦	الفجوة الرقمية وتعليم اللغة العربية الواقع والمأمول	إبراهيم صلاحالهدهد	نشر في موقع مركز الشيخ زايد لتعليم اللغة العربية لغير الناطقين بها	٢٠١٣
٤٧	اللغة العربية ومجتمع المعرفة	محمد الحناش	محاضرة في جمعية النبراس للثقافة والتنمية - وجدة - المغرب	٢٠١٣
٤٨	اللسانيات الحاسوبية العربية: الإطار والمنهج	وجدان محمد صالح كنالي	المؤتمر الدولي الثاني للغة العربية- دبي	٢٠١٣
٤٩	اللغة العربية في عصر الحوسبة والمعلوماتية: مقاربة من الهيكل العام لأنظمة العربية	صالح أبو صيني	حوليات جمعية كليات الآداب - الأردن	٢٠٠٤
٥٠	حوسبة اللغات وبعض إشكالياتها: العربية أنموذجًا	حسن كون	مجلة جيل الدراسات الأدبية و الفكرية	٢٠١٥
٥١	قراءة في الإسهامات اللسانية الحاسوبية العربية : آفاق ورهانات	فاهم سعيد	دراسات - الجزائر ع٣٦	٢٠١٥

٥٢	اللسانيات الحاسوبية مشكل المصطلح والترجمة	أحمد رضا بابا	مخبر المعالجة الآلية للغة العربية جامعة تلمسان	د.ت
٥٣	اللسانيات الحاسوبية العربية	عصام محمود	الإسكندرية، مصر : دار الوفاء لدنيا الطباعة والنشر	٢٠١٥
٥٤	المشاريع الحاسوبية على اللغة العربية والقرآن بجامعة ليدز	عبد الباقي شرف وآخرون		
٥٥	تقنيات اللغة العربية الحاسوبية معايير التقييم ورؤى التطوير	عمرو جمعة	مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية	٢٠١٦
٥٦	اللسانيات الحاسوبية: رقمنة اللغة العربية ورهان مجتمع المعرفة	إبراهيم مهديوي	شبكة الألوكة	٢٠١٦
	التحديات التي تواجه حوسبة اللغة العربية وبعض الحلول المقترحة	محمد عدني السيد وإريك اتول	الدورة العاشر للمؤتمر الدولي لعلوم وهندسة الحاسوب (ايكا ICCA)	٢٠١٦
٥٧	محاضرات في اللسانيات الحاسوبية	بن عريبة راضية	القاهرة : المجموعة العربية للتدريب والنشر،	٢٠١٧
٥٨	اللسانيات الحاسوبية في المجال الأكاديمي المغربي رفع اللبس وتحديد المفهوم	سالم الرامي	معهد الدراسات والأبحاث للتعريب بالرباط	د. ت
٥٩	اللسانيات الحاسوبية: مشكل المصطلح والترجمة	رضا بابا أحمد	مخبر المعالجة الآلية للغة العربية جامعة تلمسان)الجزائر	د. ت

	اللغويات الحاسوبية في المغرب: دراسة في الجيل الأول	عمر مهديوي	اللسانيات الحاسوبية واللغة العربية إشكالات وحلول – كنوز المعرفة – الأردن	٢٠١٨

مسرد الدراسات في المستوى النحوي:

الرقم	عنوان الدراسة	المؤلف	مكان النشر	تاريخ النشر
١	منطق النحو العربي والعلاج الحاسوبي للغات	عبد الرحمن الحاج صالح	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٢	التوليد الصوتي والنحوي والدلالي لصيغ المبني للمجهول	مازن عوض الوعر	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٣	الفعل العربي وطرق معالجته	صلاح الدين صالح حسنين	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٤	الحاسوب والنحو العربي	نبيل علي	الموسم الثقافي الرابع عشر لمجمع اللغة العربية الأردني	١٩٩٦
٥	توصيف الضمير المتصل للحاسوب : المعالجة والإشكال	مهدي أسعد عرار	أعمال المؤتمر العام للغة العربية بعنوان: قضايا الأدب واللغة والتحديات المعاصرة كلية الآداب- الجامعة الإسلامية بغزة – فلسطين	٢٠٠٠
٦	نموذج محوسب لمحلل نحوي للجمل الاسمية غير المشكولة في اللغة العربية	معتصم الحمدان	رسالة ماجستير – جامعة آل البيت الأردن	٢٠٠٢

٧	توصيف مركب العدد في اللغة العربية للحاسب الآلي	محمود الديكي	وقائع الندوة الدولية : المعالجة الآلية للغة العربية – معهد الدراسات والأبحاث للتعريب – المغرب	٢٠٠٧
٨	التركيب الإضافي في العربية: نحو توصيف جديد في ضوء اللسانيات الحاسوبية	جنات علي محمد أحمد	رسالة ماجستير – الجامعة الاردنية الأردن	٢٠٠٧
٩	توصيف النحو العربي في ضوء اللسانيات الحاسوبية : الفعل الماضي نموذجًا	أحلام عامر شريف الزبن	رسالة ماجستير – الجامعة الهاشمية الأردن	٢٠٠٨
١٠	المعالجة الآلية للغة العربية: إنشاء نموذج لساني صرفي إعرابي للفعل العربي	ارس شاشة	رسالة ماجستير – جامعة الجزائر الجزائر	٢٠٠٨
١١	توصيف نحوي للأفعال الواردة في شعر محمود درويش في ضوء اللسانيات الحاسوبية	أحمد أنيس شحادة عامر	رسالة ماجستير –الجامعة الهاشمية الأردن	٢٠٠٩
١٢	محاولة في توصيف الجملة الفعلية حاسوبيًا الملة البدوءة بالفعل الماضي التام المجرد الثلاثي الصحيح المبني للمعلوم	ريم فرحان المعايطة	المجلة الأردنية في اللغة العربية وآدابها مج ٦ ع٣	٢٠١٠
١٣	إسناد الأفعال إلى الضمائر في ضوء اللسانيات الحاسوبية	محمود مصطفى عيسى خليل	رسالة ماجستير – الجامعة الاردنية الأردن	٢٠١١

١٤	متطلبات المعاجلة الآلية للجمل الفعلية المتكلسة	عماد اللحياني	جملة بحوث جامعية كلية الآداب والعلوم الإنسانية- صفاقس – تونس	٢٠١٢
١٥	التحليل الآلي للوقف والابتداء في نصوص اللغة العربية	مجدي صوالحة وآخران	المؤتمر الدولي للحاسوب في اللغة العربية – القاهرة	٢٠١٢
١٦	طريقة تعتمد على المدونات اللغوية لتجهيز بيانات تدريب واختبار أنظمة لتجهيز الوسوم النحوية	عبدالمحسن عبيد الثبيتي وآخرون	مدينة الملك عبد العزيز للعلوم والتقنية معهد بحوث الحاسب المملكة العربية السعودية	د. ت
١٧	بناء نظام تحديد أقسام الكلام في النصوص العربية باستخدام منهجية تعلم نصف آلي	رياض سنبل وآخران	المجلة الدولية لعلوم وهندسة الحاسب باللغة العربية	د. ت
١٨	اللغة الداخلية و حوسبة البنية النحوية	سرور اللحياني	وقائع الندوة العلمية الدولية الثالثة للسانيات بعنوان اللسانيات وإعادة البناء – كلية الآداب والفنون والإنسانيات بمنوبة – تونس	٢٠١٤
١٩	توصيف الجملة الأسمية حاسوبياً (دراسة في الجملة المبدوءة بضمير المفرد المتكلم وخبرها مفرد نكرة جامد غير مضاف – للمفرد المذكر)	نبال نبيل نزال	دراسات العلوم الإنسانية والاجتماعية الأردن	٢٠١٤
٢٠	توليد الجمل العربية باستخدام لغة البرولوغ	رضا بابا أحمد	المجلة العربية لعلوم و هندسة الحاسوب، مج٦، ع١	٢٠١٥

٢١	تصنيف الأفعال والأسماء في نظرية أصناف الأشياء	عماد اللحياني وعبد الحميد عبد الواحد	مجلة اللسانيات العربية – مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية – السعودية	٢٠١٥

مسرد الدراسات في المستوى الصرفي:

الرقم	عنوان الدراسة	المؤلف	مكان النشر	تاريخ النشر
١	محلل صرفي للكلمات العربية المشتقة	آمال عبد اللطيف الرزوق	المؤتمر الثاني حول اللغويات الحسابية العربية-معهد الكويت للأبحاث العلمية	١٩٨٩
٢	التحليل الصرفي للعربية	يحيى هلال	وقائع مختارة من ندوة استخدام اللغة العربية في الحاسب الآلي- دار الرازي بيروت	١٩٨٩
٣	تمثيل الدلالة الصرفية في النظم الآلية لفهم اللغة العربية	محمد غزالي خياط	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٤	الاستكشاف الآلي للفظة الاسمية	شافية بن طامة	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٥	بحث تمثيل الدلالة الصرفية في النظم الآلية لفهم اللغة العربية	محمد غزالي خياط	(السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات) مطبوعات مكتبة الملك عبد العزيز العامة ،الرياض	١٩٩٣
٦	إحصاء الأفعال العربية في المعجم الحاسوبي	مروان البواب وآخرون	مكتبة لبنان	١٩٩٦
٧	الاصطلاح المولد GENTERM: نظام للتوليد الآلي للمصطلحات والمولدات	عبد القادر الفاسي الفهري	معهد الدراسات والأبحاث للتعريب بالرباط	١٩٩٦

٨	التحليل الصرفي للغة العربية باستخدام الحاسوب	مأمون حطاب وآخرون	الموسم الثقافي الرابع عشر لمجمع اللغة العربية الأردني	٤ أيار- ٢٢ حزيران ١٩٩٦
٩	معجم تصريف الأفعال العربية	حسن بيومي وآخرون	دار إلياس العصرية- القاهرة	
١٠	قيود تأليف الأبجديات الصرفية: لواصق تطابق الفعل والفاعل نموذجًا، في التوليد والنسقية والترجمة الآلية	عبد الرزاق تواربي	معهد الدراسات والأبحاث للتعريب بالرباط	٢٠٠١
١١	حول المولد الصرفي للكلمات المعجمية العربية، في التوليد والنسقية والترجمة الآلية	عبد الرزاق تواربي وسالم الرامي	معهد الدراسات والأبحاث للتعريب بالرباط	٢٠٠١
١٢	المنظومة الصرفية للغة العربية رؤية حاسوبية	عبد المجيد ضوة	مجمع اللغة العربية الأردني	٢٠٠٢
١٣	النظام الصرفي للعربية في ضوء اللسانيات الحاسوبية مثل من جمع التكسير	هدى سالم عبد الله آل طه	رسالة دكتوراه الجامعة الأردنية قسم اللغة العربية وآدابها	٢٠٠٥
١٤	حوسبة الصرف العربي: الموارد والخبرات اللسانية	عبد الرزاق تواربي، وخالد الأشهب، ومعد عبد الفتاح	وقائع الندوة الدولية: المعالجة الآلية للغة العربية - معهد الدراسات والأبحاث للتعريب - المغرب	٢٠٠٧
١٥	التحليل الصرفي للأسماء العربية	عبد الفتاح حمداني	مجلة أبحاث لسانية - معهد الدراسات والأبحاث للتعريب بالرباط، ع٢٣، ٢٤	٢٠٠٧

١٦	توليد الأسماء من الجذور الثلاثية الصحيحة في اللغة العربية – مقاربة لسانية حاسوبية	عمر مهديوي	جامعة الحسن الثاني – كلية الآداب والعلوم الإنسانية – الدار البيضاء	٢٠٠٨
١٧	محلل صرفي للكلمات العربية خارج السياق وداخله	عز الدين مزروعي وآخرون	اجتماع خبراء المحللات الحاسوبية الصرفية للغة العربية، دمشق	٢٠٠٩
١٨	المحللات الصرفية للغة العربية	سلوى السيد حماده	اجتماع خبراء المحللات الحاسوبية الصرفية للغة العربية دمشق	٢٠٠٩
١٩	تقرير في المحللات الصرفية للغة العربية	محمد زايد	اجتماع خبراء المحللات الحاسوبية الصرفية للغة العربية، دمشق	٢٠٠٩
٢٠	المحلل الصرفي للغة العربية لمخبر "ميراكل"	عبد المجيد بن حمادو	اجتماع خبراء المحللات الحاسوبية الصرفية للغة العربية، دمشق	٢٠٠٩
٢١	توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية	مجدي صوالحة وإيرك أتوك	اجتماع خبراء المحللات الحاسوبية الصرفية للغة العربية، دمشق	٢٠٠٩
٢٢	توصيف لغوي صرفي لشعر بدر شاكر السياب في ضوء اللسانيات الحاسوبية	عزت جهاد العجوري	رسالة ماجستير الجامعة الهاشمية – الأردن	٢٠٠٩
٢٣	التحليل الصرفي للقرآن الكريم: قاعدة بيانات مفهرسة كاملة لكامل النص القرآني	يحيى محمد الحاج وآخرون	المجلة العربية لعلوم و هندسة الحاسوب، مج٣، ع١	٢٠٠٩

٢٤	تجذير اللغة العربية باستعمال المسوقات ذات الحالات النهائية: كلمة واحدة، عدة جذور	ياسين اليونسي وآخرون	المجلة العربية لعلوم و هندسة الحاسوب، مج٣، ع١	٢٠٠٩
٢٥	استخدام النماذج الإحصائية في التعرف الآلي على خصائص المفردات العربية	يحيى محمد الحاج وآخران	المجلة العربية لعلوم و هندسة الحاسوب، مج٣، ع٢	٢٠١٠
٢٦	المقاربة الحاسوبية للصرف العربي: قراءة في الحصيلة والآفاق	عمر مهديوي	مؤتمر المحتوى العربي في الانترنت (التحديات والطموح)، مج٢ جامعة الإمام محمد بن سعود الإسلامية	٢٠١٠
٢٧	بناء نظام تحديد أقسام الكلام في النصوص العربية باستخدام منهجية تعلم نصف آلي	رياض سنبل وآخران	المجلة العربية لعلوم و هندسة الحاسوب، مج٤، ع٢	٢٠١١
٢٨	المقاربة الحاسوبية للصرف العربي : قراءة في الحصيلة والآفاق	عمرمهديوي	لسجل العلمي لمؤتمر المحتوى العربي ، صص٩٩٩-١٠٢٧.	١٤٣٢هـ ٢٠١١
٢٩	التحليل الصرفي لنصوص اللغة العربية الحديثة والكلاسيكية	مجدي صوالحة وإرك أتوك	المؤتمر الدولي للحاسوب في اللغة العربية – الرياض	٢٠١١
٣٠	معالجة الوحدات الإسمية في اللغة العربية: الأعداد المركبة نموذجاً	عزالدين غازي	أشغال الندوة الدولية(CITALA)، معهد الأبحاث والدراسات للتعريب، الرباط	٢٠١٢

٣١	بناء نظام تحديد أقسام الكلام باستخدام منهجية تعلم نصف آلي	محمد سعيد دسوقي وآخرون	المجلة الدولية لعلوم وهندسة الحاسوب باللغة العربية، العدد الثاني، المجلد الرابع، آب	٢٠١٢
٣٢	مقاربة صرفية إحصائية للتشكيل الآلي	عز الدين مزروعي وآخرون	Communications of the Arab Computer Society, Vol. 5, No. 1	٢٠١٢
٣٣	دراسة الإبدال الصرفي في ضوء اللسانيات الحاسوبية	إلهام أبو فريحة	دكتوراه جامعة العلوم الإسلامية العالمية	٢٠١٣
٣٤	تصنيف الأفعال والأسماء في نظرية أصناف الأشياء	عبد لحميد عبد الواحد	مجلة اللسانيات العربية – مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية – السعودية	٢٠١٥
٣٥	تصنيف الأفعال والأسماء في نظرية أصنافالأشياء	عماد اللحياني وعبد الحميد عبد الواحد	اللسانيات العربية مركز الملك عبد الله بن عبد العزيز لخدمة اللغة العربية– الرياض	ذو القعدة ١٤٣٦هـ – سبتمبر ٢٠١٥
٣٦	التحليل الصرفي لمكونات الكلمات العربية : دراسة لغوية حاسوبية	أحمد راغب أحمد	مجلة جيل الدراسات الأدبية و الفكرية	٢٠١٦
٣٧	لغويات حاسوبية: دراسة صوتية صرفية في أبواب الفعل الثلاثي في المعجم الوسيط	محمد جواد النوري	دار الجندي للنشر والتوزيع	٢٠١٦

٣٨	أقسام الكلم العربي نحو تقييم المحللات الصرفية العربية في ضوء منهج تمام حسان	عمرو حمدي الجندي	مجلة الكلمة، س٢٣، ع٩٢ – بيروت	٢٠١٦
٣٩	لغويات الحاسوبية، دراسة صوتية صرفية في جذور الأفعال الثلاثية	محمد جواد النوري	دار الجندي للنشر والتوزيع	٢٠١٦
٤٠	لغويات حاسوبية، دراسة صوتية صرفية في الأفعال الثلاثية المزيدة	محمد جواد النوري	دار الجندي للنشر والتوزيع	٢٠١٦

مسرد دراسات المعجم والمصطلحات:

الرقم	عنوان الدراسة	المؤلف	مكان النشر	تاريخ النشر
١	دراسة إحصائية لجذور معجم الصحاح (باستخدام الكمبيوتر)	علي حلمي موسى	الهيئة المصرية العامة لكتاب	١٩٧٨
٢	علم المصطلحات وبنوك المعطيات	ليلى المسعودي	مجلة اللسان العربي، العدد ٢٨،	١٩٨٧م
٣	المعجم الحاسوبي في نظام خبير للغة العربية	محمد مراياتي وزملاؤه	بحوث المؤتمر العلمي الأول حول الكتابة العلمية باللغة العربية – واقع وتطلعات) الذي نظمته جامعة العرب الطبية، بنغازي – ليبيا	١٩٩٠م
٤	نحو معجم عربي للتطبيقات الحاسوبية	محمود إسماعيل (الصيني)	لسجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات المنعقدة في مكتبة الملك عبدالعزز بالرياض (٨-١٢ / ١١/ ١٤١٢هـ) ، صص ٥١١-٥٢١	١٩٩٢
٥	البحث من العنوان في قواعد البيانات العربية	بخيت سليمان البخيت	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٦	تعريب المصطلحات المستعملة في الحواسيب الصغروية	أحمد بوعزي	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢

٧	في سبيل نظرية مصطلحية عربية ممكنة	محمد رشاد الحمزاوي	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٨	مصطلحات المعلوماتية واللغة العربية	سعد الحاج بكري وآخران	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٩	التخطيط لخدمات معلوماتية باللغة العربية	عبد الله الضلعان وآخران	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
١٠	بنوك المصطلحات الآلية والمعاجم الألكترونية	محمود إسماعيل صالح (الصيني)	لسجل العلمي للندوة الثانية لتعريب الحاسوب (٢٧- ٣. مارس ١٩٩٤م)	١٩٩٤م
١١	برنامج لساني-حاسوبي للتعرف الآلي على التعابير المسكوكة في اللغة العربية	محمد الحناش	مجلة التواصل اللساني ، ملحق سلسلة الندوات ، المجلد ٣	١٩٩٦
١٢	المعاجم العلمية العربية المتخصصة ودور الحاسوب	إبراهيم بن مراد	الموسم الثقافي الرابع عشر لمجمع اللغة العربية الأردني	١٩٩٦
١٣	كفاءة التحليل الصرفي في استرجاع النصوص العربية	مساعد الطيار	لة مكتبة الملك فهد الوطنية، المجلد ٤، العدد١	١٩٩٨
١٤	بنوك المصطلحات الآلية	محمود إسماعيل صيني	مجلة اللسان العربي، العدد ٤٨	١٩٩٩م
١٥	البنك الآلي السعودي للمصطلحات (باسم)	عبد الرحمن بن عبد العزيز الفاضل	مجلة اللسان العربيالعدد ٤٧	١٤٢٠هـ- ١٩٩٩

١٦	نظام حاسوبي لرومنة الأسماء العربية	عبدالملك السلمان ومنصور الغامدي وحسن الصبي	لندوة الدولية الأولى عن الحاسب واللغة العربية : الأوراق البحثية ، صص ٢١٥- ٢٢٨ .	٢٠٠٧
١٧	المتطلبات اللغوية لمعالجة التعابير الاصطلاحية العربية معالجة آلية	وفاء كامل فايد	لندوة الدولية الأولى عن الحاسب واللغة العربية : الأوراق البحثية (٢٩/ ١٠- ٢/ ١١/ ١٤٢٨هـ/ ١٠- ١٢/ ١١/ ٢٠٠٧). الرياض: مدينة الملك عبدالعزيز للعلوم والتقنية.	٢٠٠٧
١٨	الهيكلية الآلية للنصوص العربية باقتباس المفاهيم الشكلية المثالية واستعمالها للبحث في النصوص	علي الجوّة وآخرون	المجلة العربية لعلوم و هندسة الحاسوب، مج١، ع٢	٢٠٠٧
١٩	أداة ويب معتمدة على عملية التحليل الهرمي للحصول على معجم عربي موحد لتقنية المعلومات	عبدالله شرف الغامدي و بدرية سليمان الفرهود	ندوة الدولية الأولى عن الحاسب واللغة العربية : الأوراق البحثية	٢٠٠٧
٢٠	رهانات نظرية حاسوب لسانية في بناء معجم آلي للتعابير المسكوكة	عز الدين غازي	مجلة الدراسات المعجمية المغرب، ع ٦	٢٠٠٧
٢١	التعرف الموضوعي للنصوص العربية	مراد عباس وآخران	المجلة العربية لعلوم وهندسة الحاسوب، مج١، ع٣	٢٠٠٨
٢٢	التعرف الموضوعي للنصوص العربية	مراد عباس وآخران	المجلة العربية لعلوم وهندسة الحاسوب، مج٢، ع٣	٢٠٠٨

٢٣	البحث عن النصوص	علي الجوّة وآخرون	المجلة العربية لعلوم وهندسة الحاسوب، مج٢، ع٢	٢٠٠٨
٢٤	الجوانب التقييسية للمعاجم الحاسوبية	عبد المجيد بن حامدو	المنظمة العربية للتربية والثقافة والعلوم ومركز الملك عبد العزيز للعلوم والتقنية	٢٠٠٨
٢٥	الهيكلية الآلية لنتائج محركات البحث العربية والإنجليزية باقتباس المفاهيم الشكلية	علي الجوّة وآخران	المجلة العربية لعلوم وهندسة الحاسوب، مج٣، ع١	٢٠٠٩
٢٦	محركات البحث الدلالي في ظل تطبيقات الويب الدلالي	فاتن سعيد بامفلح	المجلة العربية للأرشيف والتوثيق والمعلومات. مج ١ ، ع ٢٧	٢٠١٠
٢٧	محركات البحث الدلالي في ظل تطبيقات الويب الدلالي	فاتن سعيد بامفلح	المجلة العربية للأرشيف والتوثيق والمعلومات. مج ١ ، ع ٢٧	٢٠١٠
٢٨	استخدام الشبكات العصبية في تصنيف النصوص العربية بالاستناد على طريقة تحليل القيم المفردة	فوزي حراق	المجلة العربية لعلوم وهندسة الحاسوب، مج٣، ع٢	٢٠١٠
٢٩	دراسة أدوات استخراج المصطلحات ّ النصية و مدى تكيفها مع اللغة العربية	صورية زايدي ومحمد الطيب العسكري	المجلة العربية لعلوم و هندسة الحاسوب، مج٣، ع٢	٢٠١٠
٣٠	بناء المحتوى العربي الدلالي الأول و إسهاماته في تطوير معالجة اللغة آليًا	أمل السيف وكاتجا ماركرت	المجلة العربية لعلوم و هندسة الحاسوب، مج٤، ع١	٢٠١١
٣١	نحو منهجية ومدونة في خدمة المعلوماتية العربية في الواب الاجتماعي الدلالي	إبراهيم بو نحاس ويحيى سليمان	المجلة العربية لعلوم و هندسة الحاسوب، مج٣، ع٣	٢٠١١

٣٢	نحو منهجية ومدونة في خدمة المعلوماتية العربية في الواب الاجتماعي الدلالي	إبراهيم بو نحاس ويحيى سليمان	المجلة العربية لعلوم و هندسة الحاسوب، مج٣، ع٣	٢٠١١
٣٣	بناء المحتوى العربي الدلالي الأول و اسهاماته في تطوير معالجة اللغة آليًا	أمل السيفوكاتجا ماركرت	المجلة العربية لعلوم و هندسة الحاسوب، مج٤، ع١	٢٠١١
٣٤	الاستخلاص الآلي للمحتوى العربي على شبكة الإنترنت بين الواقع والمأمول	محمد فتحي الجلاب	السجل العلمي لمؤتمر المحتوى العربي ، صص ٨٥٣-٨٩.	٢٠١١
٣٥	التحديات التي تواجه محركات البحث في استرجاع المحتوى العربي على الإنترنت: دراسة تحليلية	مؤمن النشرتي	لسجل العلمي لمؤتمر المحتوى العربي ، صص ٧١٥-٨١٤.	٢٠١١
٣٦	اقتراح خدمات الويب الدلالي كحل لاحتياجات المعمارية الموجهة للخدمات	ثابت سليمانيوالمحسن رواشد	المجلة العربية لعلوم و هندسة الحاسوب، مج٤، ع٣	٢٠١٢
٣٧	محركات البحث في النصوص العربية	مروان البواب	المؤتمر السابع لمجمع اللغة العربية بدمشق	٢٠١٢
٣٨	تطبيقات الويب الدلالي في بيئة المعرفة	علي بن ذيب الأكلبي	مجلة مكتبة الملك فهد الوطنية مج ١٨، ع٢	ذو الحجة ١٤٣٣هـ / مايو - نوفمبر ٢٠١٢م
٣٩	الحاسوب في البحث اللغوي: لسانيات المدونات اللغوية أنموذجا	محمود إسماعيل صالح	الرياض: جامعة الأميرة نورة .	٢٠١٢

٤٠	المعجم التفاعلي للغة العربية	عبد الحق الخواجة وآخران	المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية في دورته الثامنة- مصر	٢٠١٢
٤١	اقتراح خدمات الويب الدلالي كحل لاحتياجات المعمارية الموجهة للخدمات	ثابت سليماني والمحسن رواشد	المجلة العربية لعلوم و هندسة الحاسوب، مج٤، ع٣	٢٠١٢
٤٢	نحو آلية لتوليد جذاذات المعاجم العربية	حامد السحلي	ندوة معجم اللغة العربية التاريخي المركز العربي للأبحاث – الدوحة	٢٠١٢
٤٣	تطبيقات الويب الدلالي في بيئة المعرفة	علي بن ذيب الأكلبي	مجلة مكتبة الملك فهد الوطنية مج ١٨، ع٢	ذو الحجة ١٤٣٣هـ / مايو – نوفمبر ٢٠١٢م
٤٤	نحو معجم مفهرس للمصطلحات العربية الموحدة في ضوء اللسانيات الحاسوبية ومشروع الذخيرة العربية	عمر محمد أبو نواس	مجلة الدراسات اللغوية والأدبية، ع.١، ماليزيا	٢٠١٣
٤٥	نحو تأصيل منهجي لبناء أنطولوجيا اللغة العربية	مصطفى جرار	المؤتمر الدولي الثاني للغة العربية – دبي	٢٠١٣
٤٦	تطبيق العنقدة المتعددة المستويات على نص القرآن الكريم	محمد سعيد دسوقي	المجلة الدولية للتطبيقات الإسلامية في علم الحاسب والتقنية ، المجلد ٢ ،العدد ١ مارس	٢٠١٤

٤٧	المعجم الإلكتروني العربي المختص: قراءة نقدية في نماذج مختارة	أنور الجمعاوي	المؤتمر العربي الخامس لترجمة، فاس – المغرب	٢٠١٤
٤٨	أنطولوجيا الويب الدلالي ودورها في تعزيز المحتوى الرقمي: دراسة في المفاهيم والبنية الهيكلية والخدمات التفاعلية في البوابات الدلالية للتعلم الإلكتروني	أحمد فرج أحمد	مجلة المكتبات والمعلومات العربية، ع٤، مج٢٥	أكتوبر ٢٠١٥
٤٩	نحو شبكة للكلمات العربية لأغراض الصناعة المعجمية	المعتز بالله السعيد	المجلة المصرية لهندسة اللغة، ع٣	٢٠١٥
٥٠	البنك الشجري النحوي: بناؤه وتوظيفه في إطار تقنيات الذكاء الاصطناعي	أحمد روبي محمد	مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية	٢٠١٧
٥١	استخدام تقنية المعلومات للبحث في القرآن العظيم بالرسم العثماني دراسة تقييمية للمواقع القرآنية	أمير الحمامي	المجلة الدولية للتطبيقات الإسلامية في علم الحاسب والتقنية، مج٥، ع١	٢٠١٧
٥٢	منهجية فهم التعبيرات الاصطلاحية وطرق تجهيزها للمعالجة الآلية	سلوى حمادة	اللسانيات الحاسوبية واللغة العربية إشكالات وحلول – كنوز المعرفة – الأردن	٢٠١٨
٥٣	المصطلحية بين رهان المعرفية والمعالجة الرقمية	خالد اليعبودي	اللسانيات الحاسوبية واللغة العربية إشكالات وحلول – كنوز المعرفة – الأردن	٢٠١٨
٥٤	الدخلة البسيطة والدخلة المركبة من منظور لساني حاسوبي	دكيكي عبد الواحد	اللسانيات الحاسوبية واللغة العربية إشكالات وحلول – كنوز المعرفة – الأردن	٢٠١٨

مسرد دراسات الذخيرة اللغوية وما يتصل بها:

الرقم	عنوان الدراسة	المؤلف	مكان النشر	تاريخ النشر
١	استخدام الآلات الحاسبة الإلكترونية في دراسة ألفاظ القرآن الكريم	علي حلمي موسى	مجلة عالم الفكر (الكويت) المجلد ١٢،العدد ٤	١٩٨٢
٢	الذخيرة اللغوية العربية	عبد الرحمن الحاج صالح	مجلة مجمع اللغة العربية الأردني	١٩٩٨٦
٣	مسألة المصطلحات في تعريب الحاسبات	سعدي الحاج بكري	المجلة العربية للعلوم (تونس)	١٩٨٨
٤	المعالجة الآلية للكلمات وبحث النص في الأعمال المصطلحية	حسين الهبائلي	أشغال المؤتمر الرابع للسانيات: اللسانيات العربية والإعلامية- الجامعة التونسية تونس	١٩٨٩
٥	مشروع نظرية حاسوب لسانية في بناء معاجم آلية للغة العربية	محمد الحناش	مجلة التواصل اللساني، المجلد٢، العدد٢	١٩٩٠م
٦	نظرية حاسوبية لسانية لبناء المعاجم الآلية	محمد الحناش	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	١٩٩٢
٧	استخدام نظام المستشار في بناء المكانز العربية (النظم العربية المتطورة)	عبد الجبار العبد الجبار	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	١٩٩٢

٨	الحاسب الآلي وصناعة المعجم العربي	محمود فهمي حجازي	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	١٩٩٢
٩	تعريب المصطلحات المستعملة في الحواسيب الصغروية	أحمد بوعزي	السجل العلمي لندوة استخدم اللغة العربية في تقنية المعلومات) مطبوعات مكتبة الملك عبد العزيز،الرياض	١٩٩٢
١٠	نظام هبير عن اللغة العربية	سلوى أحمد الجمل	ندوة استخدام اللغة العربية في تقنية المعلومات- مكتبة الملك عبد العزيز العامة الرياض	١٩٩٣
١١	المعاجم العلمية العربية المختصة ودور الحاسوب	إبراهيم بن مراد	الموسم الثقافي الرابع عشر- مجمع اللغة العربية الأردني	١٩٩٦
١٢	المعجم الحاسوبي: إحصاء الأفعال العربية في المعجم الحاسوبي	محمد مراياتي، يحي مير علم ، محمد حسن طيان	يروت: مكتبة لبنان	١٩٩٦
١٣	الحاسوب والصناعة المعجماتية	عبد الغني أبو العزم	اللسان العربي العدد السادس والأربعون	١٩٩٨
١٤	الحاسوب والصناعة المعجماتية، اللسان العربي	عبد الغني أبو العزم	اللسان العربي العدد السادس والأربعون	١٩٩٨
١٥	البنك الآلي السعودي للمصطلحات (باسم)	عبد الرحمن بن عبد العزيز الفاضل	جلة اللسان العربي، العدد ٤٧،	١٩٩٩
١٦	المعجم العربي التأريخي الآلي	علي حلمي موسى	مجمع اللغة العربية الأردني	٢٠٠٢

١٧	الجوانب البرمجية في إعداد المعجم المفهرس للتراكيب المتشابهة لفظًا في القرآن الكريم	محمد زكي خضر	مجلة دراسات، الجامعة الأردنية، العدد ١	٢٠٠٤
١٨	رهانات نظرية حاسوبيةالسانية في بناء معجم آلي للتعابير المسكوكة	عزالدين غازي	مجلة الدراسات المعجمية المغرب، ع ٦	٢٠٠٧
١٩	استخدام ذخائر النصوص لاستخلاص المصطلحات المتخصصة	عبدالمحسن عبيد الثبيتي	لندوة الدولية الأولى عن الحاسب واللغة العربية : الأوراق البحثية ، ص ٣١-٣٨.	٢٠٠٧
٢٠	المعجم العربي في ضوء اللسانيات الحاسوبية	محمد غاليم	الندوة الدولية الأولى عن الحاسب واللغة العربية : الأوراق البحثية، صص ١٥٧ - ١٦٦.	٢٠٠٧
٢١	المعجم العربي التفاعلي التصميم المفاهيمي للمشروع	حسين إبراهيم وآخرون	مدينة الملك عبد العزيز للعلوم والتقنية والمنظمة الرياض العربية للتربية والثقافة والعلوم- تونس الرياض	٢٠٠٨
٢٢	نحو معجم عربي معاصر	محمد محمد حلمي هليل	بحوث الاجتماع الثاني لخبراء المعجم الحاسوبي للغة العربية. الرياض: مدينة الملك عبدالعزيز للعلوم والتقنية.	٢٠٠٨
٢٣	المعجم الحاسوبي للغة العربية	محمد حسن عبد العزيز، محمد يونس الحملاوي والمعتز بالله السعيد طه	لاجتماع الثاني لخبراء المعجم الحاسوبي للغة العربية المنعقد بمدينة الملك عبدالعزيز للعلوم والتقنية في الرياض	٢٠٠٨
٢٤	عمل المدونات والفهرسة الآلية	سلوى السيد حمادة	المجلة العربية لعلوم و هندسة الحاسوب، مج٢، ع٣	٢٠٠٨

٢٥	منيج إعداد المعجم العربي الحاسوبي	مروان البواب	الاجتماع الثاني لخبراء المعجم الحاسوبي التفاعلي للغة العربية- مدينة الملك عبد العزيز للعلوم، الرياض	٢٠٠٨
٢٦	معمارية المعجم العربي الالكتروني: رؤية جديدة لمعجم آلي مُبنين	عزالدين غازي ومحمدهلال	مجلة الدراسات المعجمية، المغرب، ع٨	٢٠٠٩
٢٧	استخراج ومعالجة المصطلحات: تجربة البيئة المجانية المفتوحة المصدر نوج (NooJ)	عزالدين غازي ومحمد هلال	المنتدى المصطلحي، سوسة، تونس	٢٠٠٩
٢٨	قراءة في تركيبة المعاجم الحاسوبية التفاعلية بَيْئَةُ التطوير اللغوية (NooJ) نموذجاً	عزالدين غازي	ورقة قدمت في الندوة العلمية حول "المعجمية والقاموسية والمصطلحية والمقاربات اللسانية الحديثة بجامعة منوبة"، تونس	٢٠١٠
٢٩	المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف و ترميز الأخطاء اللغوية	عبدالله يحي الفيفي	السجل العلمبي للمؤتمر الدولي لعلوم وهندسة الحاسوب في اللغة العربية في دورته الثامنة (٢٦-٢٨ ديسمبر، ٢٠١٢) جامعة القاهرة.	٢٠١٢
٣٠	التشاركية في إغناء معجم اللغة العربية التفاعلي	ندى غنيم وآخرون	مجلة جامعة دمشق للعلوم الهندسية، المجلد ٢٩ – العدد الأول	٢٠١٣

٣١	حوسبة المعجم العربي: الواقع والآفاق	جيلالي بن يشو	مجلة أرتين – المرجع الأول لطلاب الأدب. Art-En.com	٢٠١٣
٣٢	لسانيات المتون وعلوم اللغة	صالح فهد العصيمي	مجلة كلية الآداب والعلوم الإنسانية (فاس ، المغرب) العدد ١٩ ، السنة الخامسة والثلاثون	٢٠١٣
٣٣	معجم محوسب لمعاني الأفعال الثلاثية المجردة في اللغة العربية	إيمان دلول	الجامعة الإسلامية غزة (رسالة ماجستير)	٢٠١٤
٣٤	أبحاث جامعة ليدز في مجال لسانيات المدونات العربية	إيريك أتوك وعبد الله بن يحيى الفيفي	الحرف العربي والتقنية أبحاث في حوسبة العربية – مركز الملك عبد الله بن عبد العزيز الدولي – الرياض	٢٠١٥
٣٥	الذخيرة اللغوية لتراكيب القرآن الكريم	محمد زكي خضر	اللسانيات الحاسوبية واللغة العربية إشكالات وحلول – كنوز المعرفة – الأردن	٢٠١٨

٢-٣-٥-١-٤- مسرد دراسات التعرف على الحروف المكتوبة والمنطوقة:

الرقم	عنوان الدراسة	المؤلف	مكان النشر	تاريخ النشر
١	استخدام دوالّ الأرجحية في تمييز الأرقام العربية المكتوبة باليد	إيمان القيسي وحسن ناصر	مؤتمر الكويت الأول للحاسوب-جمعية الحاسب الآلي الكويتية	١٩٨٩
٢	خبرات في التعرف على الكلمات العربية المنطوقة المنفصلة	م. أ. حشيش وآخران	وقائع مختارة من ندوة استخدام اللغة العربية في الحاسب الآلي- دار الرازي بيروت	١٩٨٩
٣	معالجة اللغة العربية الطبيعية آليًا	محمد مراياتي	وقائع مختارة من ندوة استخدام اللغة العربية في الحاسب الآلي- دار الرازي بيروت	١٩٨٩
٤	تحويل نص عربي مكتوب إلى نص كتابي	ع. مرادي وآخران	وقائع مختارة من ندوة استخدام اللغة العربية في الحاسب الآلي- دار الرازي بيروت	١٩٨٩
٥	الإدراك الآلي للتضعيف	منصور محمد الغامدي	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٦	القراءة الآلية للنص العربي	حازم يوسف عبد العظيم	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢

٧	دراسة صوتية وتمييز حروف العلة الفصحى	عويزرات حاج	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٨	نظام تصحيح الهجاء	حسام الدين حسن محجوب	السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات مكتبة الملك عبد العزيز العامة الرياض	مايو ١٩٩٢
٩	المعالجة الآلية للكلام المنطوق	سالم الغزالي	ندوة استخدام اللغة العربية في المعلوماتية، المنظمة العربية للتربية والثقافة والعلوم تونس	١٩٩٦
١٠	تعامل الأجهزة والمعدات مع الحرف العربي	محمد مراياتي	ندوة استخدام اللغة العربية في المعلوماتية، المنظمة العربية للتربية والثقافة والعلوم تونس	١٩٩٦
١١	الحروف العربية والحاسوب	محمد زكي خضر	الموسم الثقافي الرابع عشر مجمع اللغة العربية الأردني	١٩٩٦
١٢	تمييز حروف اللغة العربية المكتوبة آلياً باستخدام الشبكات العصبية ذات الانتشار الرجوعي	عاصم عبد الفتاح نبوي، و صبري عبد الله محمود	مجلة جامعة الملك سعود (علوم الحاسب والمعلومات) المجلد٩	١٩٩٧م
١٣	اللسانيات الحاسوبية والترجمة الآلية	سعد عبد الستار مهدي	بيت الحكمة – بغداد	١٩٩٩
١٤	القارئ الآلي للنصوص العربية والأعداد	خضير بن بلبل	مجمع اللغة العربية الأردني	٢٠٠٢

١٥	حفص ملتقى الأصوات والحاسوب	عمرو الجندي	رسالة ماجستير جامعة القاهرة قسم اللغة والدراسات السامية والشرقية	٢٠٠٧
١٦	نظام للتعرف آلي على العناوين البريدية المخطوطة بالعربية	منصف الشرفي٨ وآخرون	المجلة العربية لعلوم و هندسة الحاسوب، ع١	٢٠٠٧
١٧	إصلاح انحناء وتقوس الكتابة في صور الوثائق العربية القديمة	منصف الشرفي وآخرون	المجلة العربية لعلوم و هندسة الحاسوب، ع٢	٢٠٠٧
١٨	مقاربة جديدة لاسترجاع التسلسل الزمني للكتابة المنفصلة باستعمال الخورزميات الجينية	عبد الكريم البعتي وآخرون	المجلة العربية لعلوم و هندسة الحاسوب، مج١، ع٣	٢٠٠٨
١٩	إصلاح انحناء وتقوس الكتابة في صور الوثائق العربية القديمة	أحمد الشريف ومحمد كحيلي	المجلة العربية لعلوم و هندسة الحاسوب، مج٢، ع٢	٢٠٠٨
٢٠	معالجة الاضطرابات التواصلية في إطار الفهم الآلي للغة العربية في الخطاب الشفوي	يونس باجو وآخرون	المجلة العربية لعلوم و هندسة الحاسوب، مج٣، ع٣	٢٠١١
٢١	القراءة الآلية للكلمات العربية المخطوطة باستعمال شبكة عصبية شفافة	سامية السنوسي المدوري	المجلة العربية لعلوم و هندسة الحاسوب، مج٤، ع٣	٢٠١٢
٢٢	تقنية التعرف الآلي على الكلام العربي – طرق جديدة لتعزيز الأداء	ضياء أبوزينة وآخرون	المجلة العربية لعلوم و هندسة الحاسوب، مج٤، ع٣	٢٠١٢

٢٣	توجيه الضوابط اللغوية والصورية للتعرف الآلي على الخط اليدوي العربي: دراسة لساني حاسوبية	سهام موساوي	رسالة دكتوراه – جامعة أبي بكر بلقايد – الجزائر	٢٠١٣
٢٤	حول نظام تمثيل الحرف العربي	مأمون حطاب	الحرف العربي والتقنية أبحاث في حوسبة العربية – مركز الملك عبد الله بن عبد العزيز الدولي – الرياض	٢٠١٥
٢٥	تقنيات التعرف الآلي على الكلام المنطوق وتطبيقاتها في القرآن الكريم: واقع وطموح	يحيى محمد الحاج	الحرف العربي والتقنية أبحاث في حوسبة العربية – مركز الملك عبد الله بن عبد العزيز الدولي – الرياض	٢٠١٥

مسرد دراسات الترجمة الآلية:

الرقم	عنوان الدراسة	المؤلف	مكان النشر	تاريخ النشر
١	الحاسب الآلي والترجمة	عبد الفتاح أبو السيدة	مجلة اللسان العربي ،العدد ٢٨	١٩٨٧
٢	الترجمة الآلية واللغة العربية	محمود إسماعيل الصيني	وقائع مختارة من ندوة استخدام اللغة العربية في الحاسب الآلي- دار الرازي بيروت	١٩٨٩
٣	بعض الصعوبات في الترجمة الآلية من الإنجليزية إلى العربية ومن العربية إلى الإنجليزية	داود عبده	المؤتمر الثاني حول اللغويات الحسابية العربية – معهد الكويت للأبحاث العلمية – الكويت	١٩٨٩
٤	الترجمة الآلية للغة العربية	محمود إسماعيل صيني	مجلة الفيصل(الرياض)،العدد ٢٣٩،	١٩٩٦
٥	التفاعل بين الإنسان والآلة في الترجمة الحاسوبية	سلمان داود الواسطي	مجلة التعريب، المركز الثقافي العربي للترجمة والتعريب، دمشق، ع٢٠	ديسمبر ٢٠٠٠
٦	مقدمة في الترجمة الآلية	عبد الله بن حمدان الحميدان	مكتبة العبيكان – الرياض	١٤٢١هـ -٢٠٠٠م
٧	الترجمة الإلكترونية: آفاق الحاضر والمستقبل	روحي بعلبكي	ضمن كتاب العربي (مستقبل الثورة الرقمية) – الكويت	٢٠٠٤
٨	أهمية التعريب في حوسبة اللغة العربية	عمر مهديوي	مجلة التعريب، المركز العربي للتعريب والترجمة والتأليف والنشر العدد ٢٦	٢٠٠٧
٩	الترجمة ا لآلية: كبنية أساسية في صرح التعريب	سلوى السيد حمادة	المجلة العربية لعلوم وهندسة الحاسوب، ع١	٢٠٠٧

١٠	الترجمة الآلية: تمثيل المعلومات لفك اللبس	سلوى السيد حماده	المجلة العربية لعلوم وهندسة الحاسوب، ع١	٢٠٠٧
١١	الترجمة الآلية ذات الصلة بالعربية من منظور تاريخي	محمد راجي زغلول وعواطف أبو الشعر، ترجمة: محمد راجي زغلول	المجلة العربية لعلوم و هندسة الحاسوب، مج٢، ع١	٢٠٠٨
١٢	في اللغة والحاسوب: الترجمة وتدقيق الإملاء بين الإنسان والآلة	داود عبده	عمان: دار جرير للنشر والتوزيع	١٤٣١هـ ٢٠١٠
١٣	حوسبة اللغة والترجمة الآلية: مسار عقدين لشركة آي تي آي	هيئة التحرير	مجلة العربية والترجمة – لبنان مج١٥، ع١٣	٢٠١٣
١٤	الترجمة الآلية من العربية وإليها		الحرف العربي والتقنية أبحاث في حوسبة العربية – مركز الملك عبد الله بن عبد العزيز الدولي – الرياض	٢٠١٥
١٥	اللسانيات الحاسوبية والترجمة الآلية	سناء منعم	إربد، الأردن : عالم الكتب نشر–توزيع– طباعة	٢٠١٥
١٦	تطويع التقنية الحديثة لخدمة الترجمة الإسلامية، نظام حرف للترجمة نموذجً	إبراهيم صالح النمي وآخران	المجلة الدولية للتطبيقات الإسلامية في علم الحاسب والتقنية، مج٤، ع٣	٢٠١٦
١٧	التقنية في خدمة الترجمة والمترجمين	محمود إسماعيل صالح	اللسانيات الحاسوبية واللغة العربية إشكالات وحلول – كنوز المعرفة – الأردن	٢٠١٨