مشروع النهوض باللغة العربية للتوجه نحو مجتمع المعرفة

PROMOTION OF THE ARABIC LANGUAGE FOR A KNOWLEDGE SOCIETY PROJECT

PROJET DE PROMOTION DE LA LANGUE ARABE POUR LA SOCIÉTÉ DU SAVOIR

المنظمة العربية للتربية والثقافة والعلوم
ALECSO
إدارة العلوم والبحث العلمي

Analyseur morphologique Al Khalil

AlKhalil Morpho Sys

برنامج الخليل الصرفي

Equipe de travail

Equipe Traitement automatique langues naturelles Laboratoire de recherche en informatique Faculté des Sciences Université Mohammed Premier Oujda Maroc	Azzeddine Mazroui (azze.mazroui@gmail.com) Abdelouafi Meziane (meziane@fso.ump.ma) Abdelhak Lakhouaja (lakhouaja@fso.ump.ma) Mohamed Ould Abdallahi Ould Bebah (medbebaha@yahoo.fr) Abderrahim Boudalal (rahimboudlal@hotmail.com) Mostafa Shoul (mshoul@hotmail.co)
Faculté des Sciences Université du Caire	Conseiller linguistique Al Moataz Bellah Al-Said (almo3tazbellah@yahoo.com)

Work Team

Team Natural Language Processing Laboratory for Computer Sc Research Faculty of Science University Mohammed First Oujda Morocco	Azzeddine Mazroui (azze.mazroui@gmail.com) Abdelouafi Meziane (meziane@fso.ump.ma) Abdelhak Lakhouaja (lakhouaja@fso.ump.ma) Mohamed Ould Abdallahi Ould Bebah (medbebaha@yahoo.fr) Abderrahim Boudalal (rahimboudlal@hotmail.com) Mostafa Shoul (mshoul@hotmail.co)
Faculty of Science Cairo University	Adviser to the language Al Moataz Bellah Al-Said (almo3tazbellah@yahoo.com)

فريق العمل

فريق المعالجة للغات الطبيعية مخبر البحث في الإعلاميات كلية العلوم جامعة محمد الأول وجدة ـ المغرب	عز الدين مزروعي azze.mazroui@gmail.com عبد الوافي مزيان meziane@fso.ump.ma عبد الحق لخواجة lakhouaja@fso.ump.ma محمد ولد عبد الله ولد بباه medbebaha@yahoo.fr عبد الرحيم بودلال rahimboudlal@hotmail.com
كلية دار العلوم جامعة القاهرة	مستشار لغوي : المعتز بالله السعيد almo3tazbellah@yahoo.com

Les coordonnateurs du projet

Dr. Amin Kalak : Coordonnateur du projet à l'ALECSO
Dr. Azzeddine Mazroui : Coordonnateur du projet au Laboratoire de Recherche en Informatique, Faculté des Sciences, Université de Mohammed I
Dr. Mansour Al-Ghamdi : Coordonnateur du projet à l'Institut de Recherche en Informatique, Ville du Roi Abdul Aziz pour la science et la technologie

Projects coordinators

Dr. Amin Kalak : Project Coordinator at ALECSO
Dr. Azzeddine Mazroui : Project Coordinator at the Reasearch Laboratory in Computer Science, Faculty of Sciences, University Mohammed First
Dr. Mansour Al-Ghamdi : Project Coordinator at the Institute for Research in Computer Sciences, King Abdul Aziz City for Sciences and Technology

منسقو المشروع

د. أمين القلق : منسق المشروع لدى المنظمة العربية للتربية والثقافة والعلوم. إدارة العلوم والبحث العلمي
د. عز الدين مزروعي : منسق المشروع لدى فريق العمل بجامعة محمد الأول كلية العلوم
د. منصور الغامدي : منسق المشروع لدى مدينة الملك عبد العزيز للعلوم والتقنية. معهد بحوث الحاسب.

Caractéristiques techniques :

- Programme Open Source
- Programme développé avec le language Java
- Utilisables avec les environnements Windows, Linux, Mac Os et Solaris
- Efficacités des interfaces interactives de l'analyseur
- Indépendance des bases de données avec le code source
- Possibilité de mettre à jour les bases de données (ajouter / supprimer / modifier)
- offrir à l'utilisateur la possibilité de spécifier la nature des sorties qu'il souhaite recevoir et le choix de la base de données utilisée dans l'analyse (la base complète des racines ou celle restreinte aux racines les plus utilisées).
- Possibilité de choisir quelques résultats d'analyse dans un fichier format CSV.
- Possibilité de recherche par la racine : quand l'utilisateur entre une racine, le programme affiche tous les mots du texte ayant cette racine comme racine possible, en plus de l'emplacement du mot dans le texte et son contexte.
- Indexation : Le programme indexe tous les mots du texte en précisant la fréquence d'apparition des mots et leur emplacement dans le texte
- Vitesse du programme : environ 10 mots par seconde pour des textes composés de dizaines de mots et 20 mots par seconde pour des textes composés de centaines de mots et 35 mots par seconde pour des textes composés de milliers de mots
- Capacité du programme : Le programme permet le traitement de textes composés de plusieurs milliers de mots.
- Enregistrement : les résultats de l'analyse peuvent être enregistrés en format HTML et CSV

Technical Characteristics :

- Open Source Program
- Program developed with Java
- Suitable for Windows, Linux, Mac OS and Solaris
- Efficient interactive interfaces
- Independence of databases from application
- Possibility to update the databases (add / delete / edit)
- Possibility to specify the desired outputs and to choice from two databases (complete database or reduced database containing the most used roots)
- Possibility to select a few test results in a CSV file
- Ability to search by the root , when the user enters a root, the program displays all the words of the text having this root, in addition to the location of the word in the text and its context.
- Indexing : the program indexes the text by specifying the frequency of occurrence of each word and its locations.
- Speed of the program : approximately 10 words per second for the texts of dozens of words and 20 words per second for the texts of hundreds of words and 35 words per second for the texts of thousands of words.
- Capacity : the program allows the processing of text consisting of several thousand of words.

أهم المزايا الفنية لبرنامج الخليل الصرفي

- برنامج مفتوح المصدر Open Source Program
- برنامج مطول بلغة البرمجة جافا Java
- إمكان استعماله في بيئة Windows و Linux و Mac Os و Solaris
- كفاءة واجهات المحلل التفاعلية مع المستعمل
- استقلالية قواعد المعطيات عن برنامج
- إمكان تحديث قواعد المعطيات (إضافة / حذف / تعديل)
- توفر البرنامج على لوحة للتحكم تسمح للمستخدم بتحديد طبيعة المخرجات التي يود الحصول عليها وقاعدة المعطيات المستعملة لذلك (اللائحة الموسعة للجذور أو لائحة الجذور الأكثر إستعمالا)
- يوفر البرنامج أيضا للمستخدم إمكانية اختيار بعض نتائج التحليل في ملف بصيغة CSV
- إمكانية البحث بالجذر : عند إدخال جذر معين يقوم البرنامج بإخراج جميع كلمات النص التي تحتوي جذورها على الجذر المدخل بالإضافة إلى موقع الكلمة داخل النص وسياقها
- الفهرسة : يقوم البرنامج بفهرسة جميع كلمات النص وذلك بتحديد عدد تردد الكلمات في النص المعالج ومواقعها داخل النص
- سرعة البرنامج : في حدود 10 كلمات في الثانية بالنسبة لنصوص من عشرات الكلمات و 20 كلمة في الثانية بالنسبة لنصوص من مئات الكلمات و 35 كلمة في الثانية بالنسبة لنصوص من آلاف الكلمات
- سعة البرنامج : يسمح البرنامج بمعالجة نصوص تتكون من آلاف الكلمات
- نسبة تغطية البرنامج : يقوم البرنامج بتحليل 92٪ من مجموع كلمات النصوص
- حفظ نتائج التحليل : يمكن حفظ نتائج التحليل بامتداد html و csv

مشروع النهوض باللغة العربية للتوجه نحو مجتمع المعرفة

PROJECT TO PROMOTE THE ARABIC LANGUAGE FOR A KNOWLEDGE SOCIETY

PROJET DE PROMOTION DE LA LANGUE ARABE POUR LA SOCIÉTÉ DU SAVOIR

المنظمة العربية للتربية والثقافة والعلوم
ALECSO
إدارة العلوم والبحث العلمي

Projet de réalisation d'un correcteur orthographique Open Source pour la langue arabe

Equipe de travail :

- Dr. Taghride Anbar / Université de Ain Shams – Égypte
  Direction générale / taghride@coltec.net
- Dr. Almoataz Billah Alsaid / Université de Caire - Égypte
  Encadrement linguistique / moataz@cu.edu.eg
- Taha Zerrouki / Université de Bouira- Algérie
  Développement informatique / taha.zerrouki@gmail.com
- Assem Chelli / Ecole Nationale supérieure d'informatique Alger - Algérie.
  Développement informatique / assem.ch@gmail.com

Coordinateur du projet à l'ALECSO

Dr. Amin Kalak / amin.kalak2012@gmail.com

Objectif du programme :

L'objectif du projet est de développer un logiciel arabe open-source, disponible sur Internet, et permettant aux utilisateurs de corriger les fautes d'orthographes qu'ils pourraient commettre lors de la rédaction d'un document en arabe.

Fonctionnalités :

-Analyse et évaluation orthographique des mots écrits.
-Identification des erreurs et proposition d'un ensemble de solutions alternatives en vue de la correction.
-Laisser à l'utilisateur l'initiative de sélectionner la bonne solution parmi les suggestions proposées.

Programme :

- Le programme est une extension du correcteur orthographique Open-Source Aspell. Le programme fournira des plugins Open-Source pour la correction de l'arabe sur le navigateur web et pour les divers projets bureautique Open-Sources.
- Le programme traite la langue arabe moderne, son programme d'apprentissage était basé sur un corpus composé de textes journalistiques, littéraires et scientifiques reflétant l'usage moderne de la langue arabe.
- Le programme assure une grande couverture linguistique, en effet, le corpus d'apprentissage contenait plus de 7 000 000 de mots. Le programme a été enrichi par une base de données de noms propres et par une base de données lexicale de l'arabe.
- Le programme propose une liste finie de termes alternatifs pour la correction. Les termes sont classés par ordre décroissant selon leurs proximités orthographiques avec le mot erroné.
- Le programme est open source, ce qui permettra à la communauté des développeurs Open-Source de prendre en main la librairie afin de la maintenir, de l'enrichir avec de nouvelles fonctionnalités et de l'intégrer dans de nouveaux usages.

Open-Source Software for Developing an Arabic Spelling Checker

Project Team :

- Prof. Taghride Anbar / University Ain Shams – Egypt
  General Direction / taghride@coltec.net
- Dr. Almoataz Billah Alsaid / University Cairo - Egypt
  language support / moataz@cu.edu.eg
- Eng. Taha Zerrouki / University Bouira - Algeria
  software development / taha.zerrouki@gmail.com
- Eng. Assem Chelli / National School of computer Alger - Algeria.
  software development / assem.ch@gmail.com

Project Coordinator at ALECSO:

Dr. Amin Kalak / amin.kalak2012@gmail.com

Project Objective

This project aims at developing an open-source Arabic Spelling Checker to be available on the Web for helping users with correcting errors that might occur when writing Arabic texts.

Software functionality

-Detecting the written word state: correct or wrong
-Suggesting the possible correct alternatives of a misspelled word
-Allowing the user to choose the suitable correct alternative

Software Desciption

-This is a plug-in software based on the open-source speller "Ayaspell". Its target is to make spelling correction available for web browsers as well as for desktop free software.
-This software deals with Modern Standard Arabic (MSA); It is based on a corpus composed of contemporary texts that belong to different domains: literature, science, media…
-Based on a corpus composed of more than 7 million words, in addition to a "Named Entities" data base, the language coverage of this system is remarkably high.
-In the correction state, a limited list of the nearest correct alternatives is suggested.
-Because this is an open-source software, developers can introduce further modifications that improve the system's functionality.

مَشرُوع تَطوير مُدَقِّق املائيّ مفتوح المَصدَر للُّغة العَرَبيَّة

فريق العمل

د. تَغريد عَنبَر (جامِعة عَين شَمس - مِصر) / الإشراف العامّ / taghride@coltec.net
د. المُعتَزّ بالله السَّعيد (جامِعة القاهِرة - مِصر) / الإطار اللُّغَويّ / moataz@cu.edu.eg
م. طه زَرُّوقِي (جامعة البويرة - الجَزائِر) / الإطار البَرمَجِيّ / taha.zerrouki@gmail.com
م. عاصِم شلي (المدرَسة الوطنيَّة العُليا للإعلام الآليّ - الجَزائِر) / الإطار البَرمَجِيّ / assem.ch@gmail.com

منسق المشروع لدى المنظمة العربية للتربية والثقافة والعلوم:

د. أمين القلق / amin.kalak2012@gmail.com

وَظيفَة البَرمَجيَّة

-مُراجعة الكلمات المكتوبة إملائيًّا، وتقييمها وفقَ معايير الصَّواب والخطأ.
-تعيين مواطن الأخطاء الإملائيَّة، وتقديم مَجموعة من البدائِل الصَّحيحة للمُستَخدِم.
-إتاحة الفُرصَة للمُستخدم لاختيار البَديل الَّذي يراهُ مُناسبًا من بين مجموعة البدائل المطروحة.

توصيف البَرمَجيَّة

-تعتمدُ البرمَجيَّةُ عَلى تَطوير لِلمُدَقِّق الإملائيّ العربيّ الحُرّ "آيسبل" لتوفير خدمة التَّدقِيق الآليّ حُرَّةً بديلةً على المُتَصَفِّحات، وعلى البرمَجيَّات المَكتَبيَّة الحرَّة وغيرها.
-تُعالِجُ البَرمَجيَّةُ اللُّغَة العربيَّة المُعاصِرة؛ إذ استُخدِمَت في تدريبها مُدَوَّنة لُغَويَّة للعربيَّة المعاصِرة، حيثُ استَمَدَّت نُصُوصَها من لُغة الصِّحافة والآداب والعُلُوم الَّتي تعكسُ الواقِعَ المعاصِرَ للغة العربيَّة.
-تُحَقِّقُ البَرمَجيَّةُ تَغطِيةً لُغَويَّةً عالية، حيثُ تجاوزَ عددُ كلمات المُدَوَّنة اللُّغَويَّة المستَخدَمة في تدريبها سبعَة ملايين كلمة؛ كما أضِيفَت قاعدةُ بياناتٍ للأعلام إلى قاعدة البيانات المُعجَميَّة للبرمجيَّة.
-تقترحُ البَرمَجيَّة قائِمةَ مُفرداتٍ صحيحةٍ ومحدودةً وقريبةً من الكلمة الخطأ مُرتَّبةً تنازليًّا بصُورةٍ قريبةٍ من المُفرَدة المُتَوَقَّعة.
-البَرمَجيَّةُ مفتوحة المَصدَر، بما يُمَكِّنُ المُطَوِّرينَ من إبداء مُقتَرَحاتٍ أو إجراء تعديلاتٍ على البَرمَجيَّة بهدف تطويرها وتحسين عملها.

مشروع النهوض باللغة العربية للتوجه نحو مجتمع المعرفة

PROJECT TO PROMOTE THE ARABIC LANGUAGE FOR A KNOWLEDGE SOCIETY

PROJET DE PROMOTION DE LA LANGUE ARABE POUR LA SOCIÉTÉ DU SAVOIR

المنظمة العربية للتربية والثقافة والعلوم
ALECSO
إدارة العلوم والبحث العلمي

مشروع تطوير مشكل آلي مفتوح المصدر للغة العربية

فريق العمل

- د. ندى غنيم - سورية، المنسق العام / nada.ghneim@gmail.com
- د. عز الدين مزروعي - المغرب/ azze.mazroui@gmail.com
- د. المعتز بالله السعيد - مصر/ moataz@cu.edu.eg
- د. غيداء ربداوي - سورية / ghaida.rebdawi@hiast.edu.sy
- م. وليد الحسن - سورية / waleed.alhasan@hiast.edu.sy
- د. سليم مصفار - تونس/ mesfarslim@yahoo.fr
- د. تغريد عنبر - مصر/ taghride@coltec.net
- د. عماد زيتوني - الولايات المتحدة الأمريكية/ imed.zitouni@gmail.com

منسق المشروع لدى المنظمة العربية للتربية والثقافة والعلوم:

د. أمين القلق / amin.kalak2012@gmail.com

هدف المشروع:

يهدف هذا المشروع إلى تطوير برمجيةٍ مفتوحةِ المصدر، تُمَكِّنُ المستخدمين من تشكيلِ النصوص المكتوبة باللغة العربية.

توصيف المشروع:

يُعَدُّ غيابُ التشكيل من التحديات التي تواجه معالجة اللغة العربية آليًا. حيث يستطيع القارئ أن يَحْزِرَ التشكيلَ الصحيح للكلمات، في حين تحتاج نظم المعالجة الحاسوبية إلى خوارزمياتٍ تحاكي قدرة البشر على استعادة التشكيل. ونظرًا لوجود تجارب تعتمد منهجيات تشكيل مختلفة (لغوية، أو إحصائية، أو هجينة)، فقد سعى فريقُ العملِ إلى مقارنة المنهجيات لإيجاد النظام الأفضل، حيث جرى:

1 - إنشاء مدوَّنةٍ لغويةٍ من نصوصٍ مشكولةٍ تشكيلاً كاملاً، تتضمن ما يزيد على 8.200.000 كلمة.

2 - تحديد معايير تقويم نظم التشكيل الآلي القابلة للقياس.

3 - بناء نظام آلي لتقويم نظم التشكيل وفق المعايير المعتمدة.

4 - تحديد نظم التشكيل التي تسمح الجهة المالكة بفتح مصدرها.

5 - تقويم نظم التشكيل المشاركة باستخدام المدونة، بهدف تحديد النظام ذي أعلى تقييم والذي سيجري اعتماده وفتح مصدره وإعداد توثيقٍ تفصيليٍّ وتقنيٍّ له.

Open-Source Software for Automatic Diacritization of Arabic Texts

Project team :

- Dr. Nada Ghneim – Syria, general coordinator / nada.ghneim@gmail.com
- Dr. Azzedine Mazroui- Morocco / azze.mazroui@gmail.com
- Dr. Almoataz Bellah Al-Said- Egypt / moataz@cu.edu.eg
- Dr. Ghaida Rebdawi – Syria / ghaida.rebdawi@hiast.edu.sy
- Eng. Waleed Al-Hasan - Syria / waleed.alhasan@hiast.edu.sy
- Dr. Slim Mesfar – Tunisia / mesfarslim@yahoo.fr
- Dr. Taghride Anbar – Egypt / taghride@coltec.net
- Dr. Imed Zitouni –USA / imed.zitouni@gmail.com

Project Coordinator at ALECSO:

Dr. Amin Kalak / amin.kalak2012@gmail.com

Project Objective:

This project aims to develop open-source software that enables users to diacritize Arabic texts.

Project Specification:

The absence of diacritization in Arabic texts is one of the challenges facing the automation of Arabic language processing. Arabic human reader can guess the words correct diacritics, whereas computing systems need algorithms that simulate the human ability to restore diacritics.

Given the existence of many experiments that use different diacritization approaches (linguistic, statistical, or hybrid), the project team sought to compare the different approaches to find the best system, by following the next steps:

1- Construction of a fully diacritized Arabic textual corpus that includes more than 8,200,000 words.

2- Identification of measurable evaluation criteria that is used to evaluate automatic diacritization systems.

3- building an automated system to evaluate the diacritization systems using adopted criteria.

4- Identification of available diacritization systems that their owners permit to open the source code.

5- Evaluating the participating diacritization systems, using the diacritized corpus, in order to identify the system with the highest evaluation, which will be adopted. The system will be available as open source, with a detailed technical documentation.

Projet de développement d'un logiciel open source libre pour la diacritization automatique des textes Arabes

Équipe de travail:

- Dr. Nada Ghneim – Syrie, Coordinateur général / nada.ghneim@gmail.com
- Dr. Azzedine Mazroui - Maroc / azze.mazroui@gmail.com
- Dr. Almoataz Bellah Al-Said- Egypt / moataz@cu.edu.eg
- Dr. Ghaida Rebdawi – Syrie / ghaida.rebdawi@hiast.edu.sy
- Eng. Waleed Al-Hasan- Syrie / waleed.alhasan@hiast.edu.sy
- Dr. Slim Mesfar – Tunisie / mesfarslim@yahoo.fr
- Dr. Taghride Anbar – Egypt / taghride@coltec.net
- Dr. Imed Zitouni –USA / imed.zitouni@gmail.com

Coordinateur du projet à l'ALECSO

Dr. Amin Kalak / amin.kalak2012@gmail.com

But du projet

Le projet vise à développer un logiciel open source qui permet aux utilisateurs la diacritization des textes écrits en langue Arabe.

Spécification du projet

L'absence de la diacritization dans les textes Arabes présente un des défis auxquels est confronté le traitement automatique de la langue Arabe. Le lecteur peut identifier la diacritization correcte des mots, tandis que les systèmes de traitement automatique ont besoin d'algorithmes qui simulent la capacité humaine pour restaurer la diacritization.

Compte tenu de l'existence de nombreuses expériences qui utilisent des approches différentes de diacritization (linguistique, statistique, ou hybride), l'équipe du projet a cherché à comparer les différentes approches afin d'identifier le meilleur système. Ainsi, elle a réalisé les étapes suivantes:

1- Construction d'un corpus linguistique entièrement diacritizé, et qui se compose de plus de 8.200.000 mots.

2- Détermination des critères d'évaluation des systèmes de diacritization automatique.

3- Développement d'un système automatisé pour l'évaluation des systèmes de diacritization selon les critères adoptés.

4- Identification des systèmes de diacritization dont le propriétaire accepte d'ouvrir leur code source.

5- Évaluation des systèmes de diacritization participants, en utilisant le corpus, afin d'identifier le système ayant la meilleure évaluation pour l'adopter et le fournir en tant que logiciel open source, et la préparation d'une documentation technique détaillée de ce système.

مشروع النهوض باللغة العربية للتوجه نحو مجتمع المعرفة

PROJECT TO PROMOTE THE ARABIC LANGUAGE FOR A KNOWLEDGE SOCIETY

PROJET DE PROMOTION DE LA LANGUE ARABE POUR LA SOCIÉTÉ DU SAVOIR

المنظمة العربية للتربية والثقافة والعلوم
ALECSO
إدارة العلوم والبحث العلمي

SEWAC : Wordnet Sémantique de l'Arabe

Equipe de travail

• Dr. Lamia Hadrich Belguith, (ANLP Research group,MIRACL, University of Sfax)
Coordinateur général du projet / l.belguith@fsegs.rnu.tn
• Dr. Abdelmajid Ben Hamadou, Coordinateur de l'équipe tunisienne
abdelmajid.benhamadou@isimsf.rnu.tn
• Dr. Nada Ghenim, Coordinateur de l'équipe syrienne
nada.ghneim@gmail.com
• Dr. Almotaz bellah Al-said, Coordinateur de l'équipe egyptienne
moataz@cu.edu.eg
• Dr. Mohamed Hannach, Coordinateur de l'équipe marocaine
elhannach@yahoo.com

Coordinateur du projet à l'ALECSO ; Dr. Amin Kalak
amin.kalak2012@gmail.com

Informations

Réalisation d'une plateforme pour l'analyse syntaxique de textes arabes. Cette plateforme comporte plusieurs outils d'analyse : partant de la segmentation des textes arabes en phrases et allant à la reconnaissance des unités structurelles et fonctionnelles de chaque phrase et la détermination des relations les reliant. Ces outils sont les suivants :

- Un outil pour la segmentation du textes en paragraphes, phrases et unités lexicales en se basant sur un ensemble de règles qui étudient le contexte des conjonctions de coordination et d'autres mots outils.
- Un outil pour l'analyse morphologique des mots permettant de déterminer toutes les caractéristiques morphologiques possibles pour chaque mot afin d'identifier sa catégorie grammaticale (nom, verbe, adjectif, etc.) et d'autres caractéristiques comme le nombre (singulier, duel, pluriel), le genre (masculin, féminin), la détermination (déterminé, non déterminé), le cas (nominatif, accusatif, génitif), le temps, la personne pour les verbes, etc.
- Un outil pour la reconnaissance des entités nommées (Named entity recognition) permettant de reconnaître les noms de personnes, d'organisations, de pays, etc.
- Un outil pour l'analyse syntaxique basé sur l'apprentissage automatique et permettant la reconnaissance des unités principales de la phrase et des relations structurelles et fonctionnelles entre elles.

De plus, cette plateforme offre un corpus annoté créé dans le cadre de ce projet pour l'apprentissage et l'évaluation du système d'analyse syntaxique. Ce corpus reflète l'arabe moderne et comporte plus que 160 milles mots.

SEWAC : Semantic Wordnet of Arabic

Work Team

• Dr. Lamia Hadrich Belguith, (ANLP Research group,MIRACL, University de Sfax)
General Project Coordinator / l.belguith@fsegs.rnu.tn
• Dr. Abdelmajid Ben Hamadou, Tunisian team Coordinator
abdelmajid.benhamadou@isimsf.rnu.tn
• Dr. Nada Ghenim, Syrian team Coordinator
nada.ghneim@gmail.com
• Dr. Almotaz bellah Al-said, Egyptian team coordinator
moataz@cu.edu.eg
• Dr. Mohamed Hannach, Moroccan team Coordinator
elhannach@yahoo.com

Project Coordinator at ALECSO ; Dr. Amin Kalak
amin.kalak2012@gmail.com

Project goal

Realization of an Arabic text parsing platform. This platform includes several tools: from the text segmentation into sentences to the recognition of the structural and functional sentence units and the identification of the relationships between them. These tools are the following :

- A tool to segment texts into paragraphs, sentences and lexical units based on a rule set that studies the context of the coordination conjunctions and other tool words.
- A tool for morphological analysis that determines for each word all its possible morphological features in order to identify its part of speech (noun, verb, adjective, etc.) and other features such as the number (singular, dual, plural), the gender (masculine, feminine), the determination (determined, not determined), the case (nominative, accusative, genitive), the tense, the personal pronoun for the verbs, etc.
- A tool for the named entity recognition to identify the names of persons, organizations, countries, etc..
- A parsing tool based on machine learning and allowing the identification of the main sentence units and their structural and functional relations.

In addition, this platform offers an annotated corpus constructed especially for this project and used for the parser learning and evaluation. This corpus reflects modern Arabic and involves more than 160 thousands words.

سواك : الشبكة الدلالية للعربية

فريق العمل

• د. لمياء هدريش بلغيث ، المشرف العام، مخبر ميراكل جامعة صفاقس / l.belguith@fsegs.rnu.tn
• د.عبدالمجيد بن حمادو ، منسق الفريق التونسي
abdelmajid.benhamadou@isimsf.rnu.tn
• د.ندى غنيم ، منسق الفريق السوري / nada.ghneim@gmail.com
• د.المعتز بالله السعيد ، منسق الفريق المصري / moataz@cu.edu.eg
• د.محمد الحناش منسق الفريق المغربي / elhannach@yahoo.com

منسق المشروع لدى المنظمة العربية للتربية والثقافة والعلوم
د. أمين القلق
amin.kalak2012@gmail.com

الهدف من المشروع :

إنجاز منصة platform للتحليل النحوي للنصوص العربية تشتمل على عدة أنظمة مساعدة للتحليل انطلاقا من تقسيم النص الى جمل ووصولا إلى التعرف على عناصر كل جملة وتحديد العلاقات التركيبية والوظيفية التي تربط بينها .وتتمثل هذه الأنظمة في التالي :

أولا- نظام لتقسيم النص إلى فقرات و جمل ومفردات باعتماد مجموعة من القواعد التي تعتمد على سياق حروف العطف و بعض الأدوات الأخرى

ثانيا- نظام للتحليل الصرفي للكلمات يمكن من إسناد الخصائص الصّرفيّة الممكنة لكلّ كلمة قصد التعرف على صفتها النّحويّة (اسم، فعل، صفة، أداة ،الخ) و خصائص أخرى مثل العدد (مفرد، مثنى، جمع)، الجنس (مؤنّث، مذكر) ، التعريف (معرفة، نكرة) ، حالة الإعراب (رفع، نصب، جزم) ، الزّمان و الضمير بالنسبة للأفعال ، الخ

ثالثا - نظام للتعرف على المكونات الإسمية (Named Entity Recognition)
يمكن من التعرف على أسماء الأشخاص والمؤسسات والبلدان الخ.

رابعا - نظام للتحليل النحوي يعتمد على التلقين الآلي ويمكن من التعرف على عناصر الجملة والعلاقات التركيبية والوظيفية التي تربط بينها.

كما تشمل هذه المنصة (platform) على مدونة مذيلة (Annotated corpus) أنجزت خصيصا لهذا المشروع في نطاق تلقين وتقييم نظام للتحليل النحوي.

تعكس هذه المُدَوَّنة واقِعَ اللُّغة العربيَّة المُعاصرة و تحتوي على أكثر من 160 ألف كلمة.

www.alecso.org.tn

مشروع النهوض باللغة العربية للتوجه نحو مجتمع المعرفة

PROJECT TO PROMOTE THE ARABIC LANGUAGE FOR A KNOWLEDGE SOCIETY

PROJET DE PROMOTION DE LA LANGUE ARABE POUR LA SOCIÉTÉ DU SAVOIR

المنظمة العربية للتربية والثقافة والعلوم
ALECSO
إدارة العلوم والبحث العلمي

Une platforme pour l'analyse syntaxique de textes arabes

Equipe de travail

- Dr. Lamia Hadrich Belguith, (ANLP Research group,MIRACL, University of Sfax) Coordinateur général du projet / l.belguith@fsegs.rnu.tn
- Dr. Abdelmajid Ben Hamadou, Coordinateur de l'équipe tunisienne abdelmajid.benhamadou@isimsf.rnu.tn
- Dr. Nada Ghenim, Coordinateur de l'équipe syrienne nada.ghneim@gmail.com
- Dr. Almotaz bellah Al-said, Coordinateur de l'équipe egyptienne moataz@cu.edu.eg
- Dr. Mohamed Hannach, Coordinateur de l'équipe marocaine elhannach@yahoo.com

Coordinateur du projet à l'ALECSO ; Dr. Amin Kalak
amin.kalak2012@gmail.com

Informations

Réalisation d'une plateforme pour l'analyse syntaxique de textes arabes. Cette plateforme comporte plusieurs outils d'analyse : partant de la segmentation des textes arabes en phrases et allant à la reconnaissance des unités structurelles et fonctionnelles de chaque phrase et la détermination des relations les reliant. Ces outils sont les suivants :

- Un outil pour la segmentation du textes en paragraphes, phrases et unités lexicales en se basant sur un ensemble de règles qui étudient le contexte des conjonctions de coordination et d'autres mots outils.
- Un outil pour l'analyse morphologique des mots permettant de déterminer toutes les caractéristiques morphologiques possibles pour chaque mot afin d'identifier sa catégorie grammaticale (nom, verbe, adjectif, etc.) et d'autres caractéristiques comme le nombre (singulier, duel, pluriel), le genre (masculin, féminin), la détermination (déterminé, non déterminé), le cas (nominatif, accusatif, génitif), le temps, la personne pour les verbes, etc.
- Un outil pour la reconnaissance des entités nommées (Named entity recognition) permettant de reconnaître les noms de personnes, d'organisations, de pays, etc.
- Un outil pour l'analyse syntaxique basé sur l'apprentissage automatique et permettant la reconnaissance des unités principales de la phrase et des relations structurelles et fonctionnelles entre elles.

De plus, cette plateforme offre un corpus annoté créé dans le cadre de ce projet pour l'apprentissage et l'évaluation du système d'analyse syntaxique. Ce corpus reflète l'arabe moderne et comporte plus que 160 milles mots.

Platform for parsing Arabic Texts

Work Team

- Dr. Lamia Hadrich Belguith, (ANLP Research group,MIRACL, University de Sfax) General Project Coordinator / l.belguith@fsegs.rnu.tn
- Dr. Abdelmajid Ben Hamadou, Tunisian team Coordinator abdelmajid.benhamadou@isimsf.rnu.tn
- Dr. Nada Ghenim, Syrian team Coordinator nada.ghneim@gmail.com
- Dr. Almotaz bellah Al-said, Egyptian team coordinator moataz@cu.edu.eg
- Dr. Mohamed Hannach, Moroccan team Coordinator elhannach@yahoo.com

Project Coordinator at ALECSO ; Dr. Amin Kalak
amin.kalak2012@gmail.com

Project goal

Realization of an Arabic text parsing platform. This platform includes several tools: from the text segmentation into sentences to the recognition of the structural and functional sentence units and the identification of the relationships between them. These tools are the following :

- A tool to segment texts into paragraphs, sentences and lexical units based on a rule set that studies the context of the coordination conjunctions and other tool words.
- A tool for morphological analysis that determines for each word all its possible morphological features in order to identify its part of speech (noun, verb, adjective, etc.) and other features such as the number (singular, dual, plural), the gender (masculine, feminine), the determination (determined, not determined), the case (nominative, accusative, genitive), the tense, the personal pronoun for the verbs, etc.
- A tool for the named entity recognition to identify the names of persons, organizations, countries, etc..
- A parsing tool based on machine learning and allowing the identification of the main sentence units and their structural and functional relations.

In addition, this platform offers an annotated corpus constructed especially for this project and used for the parser learning and evaluation. This corpus reflects modern Arabic and involves more than 160 thousands words.

منصة لتحليل النصوص العربية

فريق العمل

- د. لمياء هدريش بلغيث ، المشرف العام، مخبر ميراكل جامعة صفاقس / l.belguith@fsegs.rnu.tn
- د.عبدالمجيد بن حمادو ، منسق الفريق التونسي abdelmajid.benhamadou@isimsf.rnu.tn
- د.ندى غنيم ، منسق الفريق السوري / nada.ghneim@gmail.com
- د.المعتز بالله السعيد ، منسق الفريق المصري / moataz@cu.edu.eg
- د.محمد الحناش منسق الفريق المغربي / elhannach@yahoo.com

منسق المشروع لدى المنظمة العربية للتربية والثقافة والعلوم
د. أمين القلق
amin.kalak2012@gmail.com

الهدف من المشروع :

إنجاز منصة platform للتحليل النحوي للنصوص العربية تشتمل على عدة أنظمة مساعدة للتحليل انطلاقا من تقسيم النص الى جمل ووصولا إلى التعرف على عناصر كل جملة وتحديد العلاقات التركيبية والوظيفية التي تربط بينها .وتتمثل هذه الأنظمة في التالي :

أولا- نظام لتقسيم النص إلى فقرات و جمل ومفردات باعتماد مجموعة من القواعد التي تعتمد على سياق حروف العطف و بعض الأدوات الأخرى

ثانيا- نظام للتحليل الصرفي للكلمات يمكن من إسناد الخصائص الصّرفيّة الممكنة لكلّ كلمة قصد التعرف على صفتها النّحويّة (اسم، فعل، صفة، أداة ،الخ) و خصائص أخرى مثل العدد (مفرد، مثنى، جمع)، الجنس (مؤنّث، مذكر) ، التعريف (معرفة، نكرة) ، حالة الإعراب (رفع، نصب، جزم) ، الزّمان و الضمير بالنسبة للأفعال ، الخ

ثالثا - نظام للتعرف على المكونات الإسمية (Named Entity Recognition)

يمكن من التعرف على أسماء الأشخاص والمؤسسات والبلدان الخ.

رابعا - نظام للتحليل النحوي يعتمد على التلقين الآلي ويمكن من التعرف على عناصر الجملة والعلاقات التركيبية والوظيفية التي تربط بينها.

كما تشمل هذه المنصة (platform) على مدونة مذيلة (Annotated corpus) أنجزت خصيصا لهذا المشروع في نطاق تلقين وتقييم نظام للتحليل النحوي.

تعكس هذه المُدَوَّنة واقِعَ اللُّغة العربيَّة المُعاصرة و تحتوي على أكثر من 160 ألف كلمة.