محاور الدورة
الثانية

صناعة اللغات
Languages Industry
حلول لغويةً واعدة

المدونات الحاسوبية : نافذة الإبداع في الترجمة

دورة

د.سلطان المجيول

٦ ساعات بتاريخ ١ و ٢ نوفمبر

من ٨ - ١١ مساءً
من يوم الدورة عبر

zoom

أستاذ المدونات اللغوية واللغويات الحاسوبية المشارك
بجامعة الملك سعود

محاور الورشة

- مفهوم المدونات المتعلقة بالترجمة وهيكلة بياناتها
- أيُّ مدونة ترجمية لأي مهنة ترجمية؟
- من التحذية الآلية إلى التحذية المشاهدة:
أهم الاستنتاجات التطبيقية
- استخدام المدونات الترجمية للتعرف على طبيعة اللغة بين لغتين:
نموذج SMT

SCAN ME

محاور التدريب لليوم الأول: كيف تتم التحذية؟

- هيكلة البيانات لأنواع المدونات الترجمية
- أي مدونة ترجمية لأي مهنة ترجمية
- التحذية والتطبيق في SkE

محاور التدريب لليوم الثاني

- نماذج إحصائية مدونية
- SMT
- NMT

محاور التدريب لليوم الأول: كيف تتم التحذية؟

- هيكلة البيانات (كلمة بكلمة/ عبارة بعبارة/ جملة بجملة/ فقرة بفقرة)
- أي مدونة ترجمية لأي مهنة ترجمية (الوعاء genre/ المجال domain/ الموضوع topic)

http://www.glottopedia.org/index.php/Parallel_corpus

http://www.ilc.cnr.it/EAGLES96/corpustyp/node1.html

https://www.clarin.eu/resource-families/parallel-corpora

https://www.ldc.upenn.edu/

البيانات والمحاذاة Data and Alignment

هناك عدة إجراءات لإعداد بيانات النصوص بين اللغتين أو أكثر. هنا سنقدم هذه الطرق من الأبسط إلى الأكثر تقدما

عدة أدوات تعمل على تحذية النصوص بين لغتين، وذلك بوضع جملة إزاء جملة:

1. FarkasAndras: open-source aligner
2. AlignFactory light
3. 1Plus Tools
4. TextAlign
5. AntPConc

5 CAT tools

1. TRANSLATION MEMORY SOFTWARE

Trados Workbench, DéjàVuX, SDLX, Star Transit, MultiTrans, Similis, MetaTexis

2. LANGUAGE SEARCH-ENGINE SOFTWARE
3. TERMINOLOGY MANAGEMENT SOFTWARE

SDL MultiTerm, LogiTerm and Termex

4. ALIGNMENT SOFTWARE

Bitext2, Tmx Bligner, YouAlign and LF Aligner

5. INTERACTIVE MACHINE TRANSLATION

I have used Bitext2: https://sourceforge.net/projects/bitext2tmx/files/latest/download

Sketch Engine (powerful web-based tool)

Sketch Engine (powerful web-based tool)

- mono-lingual
  1. **Word Sketch**: collocations and word combinations
  2. **Word Sketch Difference**: Compare collocations of two words
- Lexical semantics
  3. **Thesaurus**: Synonyms and similar words
- bi-multi-lingual
  4. **Concordance**: Examples of use in context
  5. **Parallel Concordance**: Translation search
- **corpus applications**
  6. **Wordlist**: Frequency list
  7. **N-grams**: Multiword expressions
  8. **Keywords**: Terminology extraction
  9. **Text type analysis**: Statistics of the corpus
  10. **OneClick Dictionary**: Automatic dictionary drafting

Sketch Engine

Word Sketch Difference: Compare collocations of two words

أولا: إضافة مدونة بصيغة txt أو csv أو xls أو xlsx أو TMX أو XML أو XLIFF

ثانيا: يتضمن الملف عموديْن؛ الأول: اللغة المصدر، والثانية: اللغة الهدف، وفي كل صف من صفوف العمودين الجمل المتوازية بين اللغتين.

> Add English corpus
> Add Arabic corpus
> Compile bilingual corpora

Sketch Engine

Parallel Concordance

search parallel corpus

To search the corpus as a parallel corpus, first select the corpus in the language that should appear on the left and then, when setting the search criteria, select the other language(s). Multiple languages can be selected to display a multilingual concordance.

المدونات الحاسوبية نافذة الإبداع في الترجمة

محاور التدريب لليوم الثاني

- إحصاءات SkE
- نموذج SMT في مقابل NMT

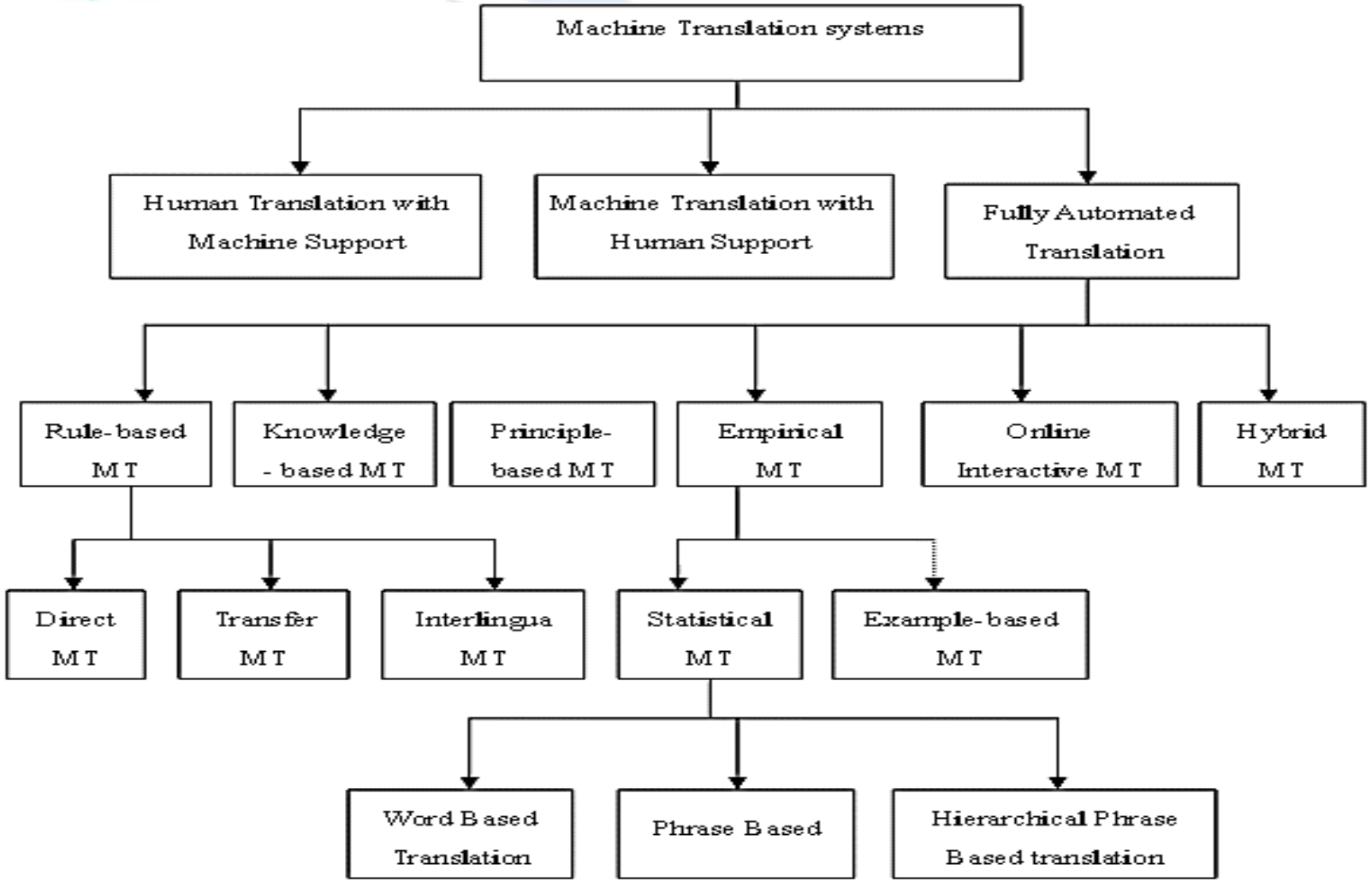
✓ Statistics in Sketch Engine
✓ Statistical (Machine Translation)
✓ Neural Machine Translation System (Deep Learning)

المدونات الحاسوبية نافذة الإبداع في الترجمة

مسح أبحاث الترجمة الآلية

http://www.statmt.org/survey/Topic/WordBasedModels

أنظمة الترجمة الآلية



Statistical Machine Translation

- تعمل نماذج SMT على تكشيف التمثيل بين لغتين وفق 5 مراحل تُعرف بـ:
- **IBM MODEL 1**
- **IBM MODEL 2**
- **IBM MODEL 3**
- **IBM MODEL 4**
- **IBM MODEL 5**
- تتلخص فكرة هذه النموذج في معالجة المساحة **SPACE** بين تمثيل الملف (الهدف) وتمثيل الاستعلام (المصدر):

- **Mapping the document representation into the query representation space: document translation**
- **Mapping the Query representation into the document representation space: Query translation**
- **Mapping the both document and query representations into third space.**

Statistical Machine Translation

- Statistical MT relies on translation examples contained in a parallel corpus, i.e., a set of texts translated into another language.
- Such a corpus can be further processed into aligned sentences
- Word-to-word translation
- Phrase-based translation
- Syntax-based translation (tree-to-tree, tree-to-string)
- Trained on parallel corpora
- Usually noisy-channel (at least in spirit)

IBM Model 1

ترجمة الكلمة وفق المعاجم

طريق: way, road, path, track, course, route

وفق السياقات المتعددة قد نجد أن كلمة way وroad أكثر الكلمات تواردًا مع (طريق)

في معالجة الترجمة وفق IBM Model 1

لنفترض الآتي:

$$p_f(e) = \begin{cases} 0.65 \ if \ e = way \\ 0.20 \ if \ e = road \\ 0.02 = if \ e = path \\ 0.02 \ if \ e = track \\ 0.10 \ if \ e = course \\ 0.01 \ if \ e = route \end{cases}$$

ترجمة (طريق)	التكرار
way	65
road	20
path	2
track	2
course	10
route	1

IBM Model 1

ترتيب الترجمة بين الإنجليزية والعربية تشكل تحدٍ كبير.

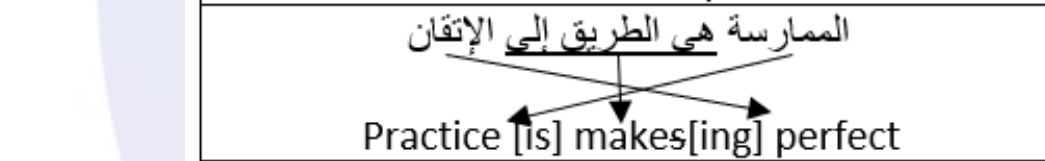
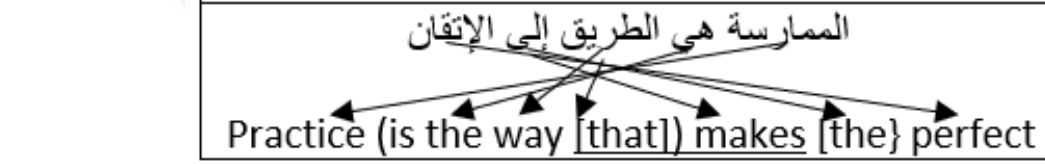
في نموذج الـ IBM Model 1 أربع حالات محتملة في الترتيب بين العربية والإنجليزية:

تراتب متطابق	أنا آخذ ال طريق I take your way
تراتب معاد ترتيبه	أ هذا هو ال طريق إلى ال محطة ؟ Is this the way to the station ?
تراتب ناقص	أ هذا هو ال طريق إلى ال محطة ؟ Is this the way to the station ?
تراتب زائد	الممارسة هي الطريق إلى الإتقان Practice makes perfect

> عُولجت التراتب المعاد ترتيبها في IBM Model 2 ثم عوجلت مشكلات التراتب الناقص والزائد بنماذج 3 و4.
> ظهرت مشكلات في نماذج 3 و4 ثم بُني نموذج IBM Model 5.

IBM Models 3 and 4

> أُجري نموج التخصيب **fertility** لتحسين مشكلات الاحتمالات التوزيعية على مستوى الكلمة–الكلمة والعبارة–العبارة (نموذج **2**)

	fertility step	مرحلة التخصيب
	NULL insertion step	مرحلة إدخال الصفر
	lexical translation step	مرحلة الترجمة المعجمية
	distortion step	مرحلة التوزيع

> أُجرى نموذج إعادة الترتيب من النموذج **2** في النموذج **3** ليظهر نموذج **4** القائم على الاحتمالات النسبية من النماذج السابقة

IBM Models 1–4 are *deficient*

IBM Model 5

حاول هذا النموذج معالجة القصور بـ:

> تضمين الترجمات الخاطئة في النص كاحتمال إيجابي (دلالة على الخطأ)

> محاولة الكشف عن ذلك ببيانات تدريب واختبار ضخمة

> الوصول إلى أن نماذج مثل هذه لن تنجح، وأن الاتجاه إلى بيانات ضخمة بالتعلم العميق للترجمة الآلية هو الحل الدائم في المستقبل

حلولٌ لغوية واعدة

Deep Learning for Machine Translation

Bilingual Evaluation Understudy BLEU

It assigns a higher value to outputs which have a larger number of matching n-grams with some given references. At the same time it aims to guarantee some balance between the length of translation and the reference sentences

صناعة اللغات
Languages Industry

شكراً لكم

د. سلطان المجيول