A proposal for an Arabic-to-English SMT system

Cristina España i Bonet

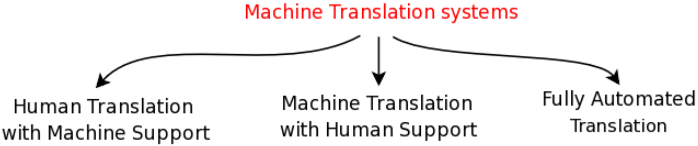
Advisor: Dr. Lluís Màrquez Villodre

22th February, 2008

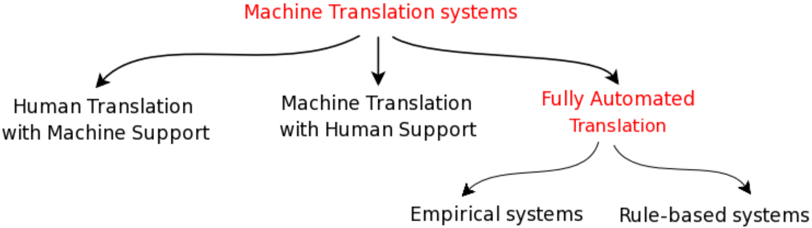
Overview

Goal: *Apply Machine Translation techniques to translate from Arabic to English in the context of the 2008 NIST Machine Translation Open evaluation*

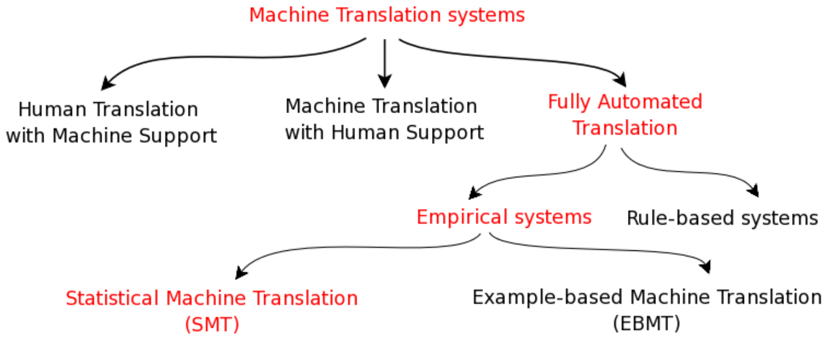
Goal: *Apply Machine Translation techniques to translate from Arabic to English in the context of the 2008 NIST Machine Translation Open evaluation*



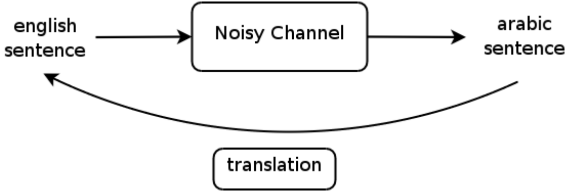
Goal: *Apply Machine Translation techniques to translate from Arabic to English in the context of the 2008 NIST Machine Translation Open evaluation*



Goal: *Apply Machine Translation techniques to translate from Arabic to English in the context of the 2008 NIST Machine Translation Open evaluation*



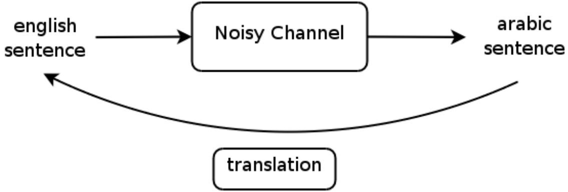
Basics:


english
sentence
Noisy Channel
arabic
sentence
translation

Statistical Machine Translation

Translation as a Noisy Channel Model

Basics:



Mathematically:

$$P(e|f) = \frac{P(e)\,P(f|e)}{P(f)}$$

$$T(f) = \hat{e} = \mathrm{argmax_e}\ P(e|f) = \mathrm{argmax_e}\ P(e)\,P(f|e)$$

$P(e)$ Language Model

- Probability scores of n-grams in the target language
- Takes care of correctness and fluency
- Data: corpora in the target language

$P(e)$ Language Model

- Probability scores of n-grams in the target language
- Takes care of correctness and fluency
- Data: corpora in the target language

$P(f|e)$ Translation Model

- IBM models
- Data: aligned corpora in source and target languages

Statistical Machine Translation

Components

$P(e)$ Language Model

- Probability scores of n-grams in the target language
- Takes care of correctness and fluency
- Data: corpora in the target language

$P(f|e)$ Translation Model

- IBM models
- Data: aligned corpora in source and target languages

argmax

- Search done by the *decoder*

Maximum Likelihood estimate $\leftrightarrow$ Maximum Entropy estimate

Maximum Likelihood estimate $\leftrightarrow$ Maximum Entropy estimate

$$\Downarrow$$

$$\mathrm{argmax}\ \log P(e|f) = \mathrm{argmax}\ \sum_m \lambda_m h_m(f|e)$$

Maximum Likelihood estimate $\leftrightarrow$ Maximum Entropy estimate

$$\Downarrow$$

$$\mathrm{argmax}\ \log P(e|f) = \mathrm{argmax}\ \sum_m \lambda_m h_m(f|e)$$

- $h_m \longrightarrow$ features (log-probabilities)
  - Language and translation models,
  - and distortion, word penalty, phrase penalty...

- $\lambda_m \longrightarrow$ weight of every feature

Maximum Likelihood estimate $\leftrightarrow$ Maximum Entropy estimate

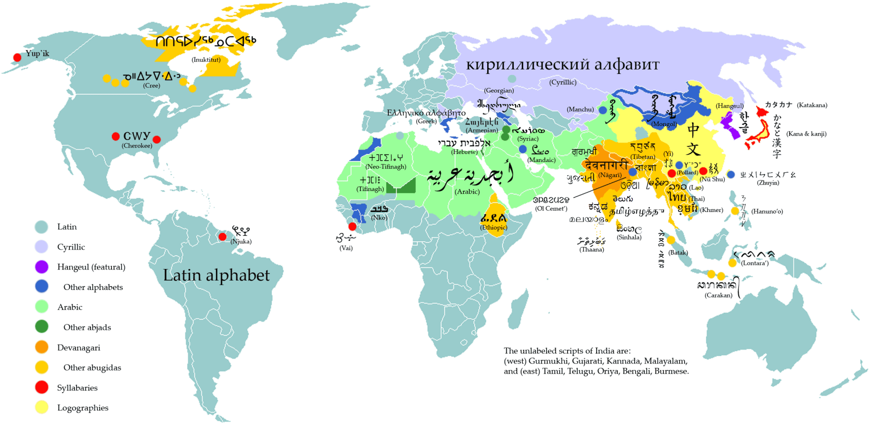
$$\Downarrow$$

$$\mathrm{argmax}\ \log P(e|f) = \mathrm{argmax}\ \sum_m \lambda_m h_m(f|e)$$

- $h_m \longrightarrow$ features (log-probabilities)
  - Language and translation models,
  - and distortion, word penalty, phrase penalty...

☞ $\lambda_m \longrightarrow$ weight of every feature

Language pair
Arabic-English



Language pair
Arabic

وتأتي دول فرنسا وبريطانيا وايطاليا وألمانيا وأيرلندا وأسبانيا ولو كسمبورج في المقدمة
وبخلاف الشركات الأوروبية فقد وصل حجم رءوس الأموال المصدرة لل شركات العاملة في
مصر حتي ديسمبر ٢٠٠٠ الي ١٢٦ مليار ج نيه لعدد ١٠ آلاف شركة ا ستثماري ة
مؤسسة وفقا لقانون الاستثمار ..

☞ Right to left text (numerals: left to right)

- VSO structure
- Alphabet: allographic variants, diacritics and ligatures
- Agglutinative language

Language pair
Arabic

وتأتي دول فرنسا وبريطانيا وايطاليا وألمانيا وأيرلندا وأسبانيا ولو كسمبورج في المقدمة
وبخلاف الشركات الأوروبية فقد وصل حجم رءوس الأموال المصدرة لل شركات العاملة في
مصر حتي ديسمبر ٢٠٠٠ الي ١٢٦ مليار ج نيه لعدد ١٠ آلاف شركة ا ستثماري ة
مؤسسة وفقا لقانون الاستثمار .

- Right to left text (numerals: left to right)
- ☞ VSO structure
- Alphabet: allographic variants, diacritics and ligatures
- Agglutinative language

Language pair
Arabic

وتأتي دول فرنسا وبريطانيا وايطاليا وألمانيا وأيرلندا وأسبانيا ولو كسمبورج في المقدمة
وبخلاف الشركات الأوروبية فقد وصل حجم رءوس الأموال المصدرة لل شركات العاملة في
مصر حتي ديسمبر ٢٠٠٠ الي ١٢٦ مليار ج نيه لعدد ١٠ آلاف شركة ا ستثماري ة
مؤسسة وفقا لقانون الاستثمار ..

- Right to left text (numerals: left to right)
- VSO structure
- ☞ Alphabet: allographic variants, diacritics and ligatures
- Agglutinative language

Language pair
Arabic

وتأتي دول فرنسا وبريطانيا وايطاليا وألمانيا وأيرلندا وأسبانيا ولو كسمبورج في المقدمة
وبخلاف الشركات الأوروبية فقد وصل حجم رءوس الأموال المصدرة لل شركات العاملة في
مصر حتي ديسمبر ٢٠٠٠ الي ١٢٦ مليار ج نيه لعدد ١٠ آلاف شركة ا ستثماري ة
مؤسسة وفقا لقانون الاستثمار ..

- Right to left text (numerals: left to right)
- VSO structure
- Alphabet: allographic variants, diacritics and ligatures
- ☞ Agglutinative language

Language pair
Arabic: diacritics

What are diacritics?

1. Short vowels: *fatha*, *kasra* and *damma*
2. Non-vowel mark: *sukun*
3. Double consonant mark: *shadda*

What are diacritics?

1. Short vowels: *fatha*, *kasra* and *damma*
2. Non-vowel mark: *sukun*
3. Double consonant mark: *shadda*

☞ But diacritics are not usually seen in written texts: MT corpora are non-vocalized and non-diacritized.

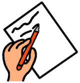
Language pair

Arabic: diacritics and ambiguity

The absence of diacritics increases the ambiguity

ktb

katab



to write

katib

writer

kitab

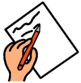
book

Language pair
Arabic: diacritics and ambiguity

The absence of diacritics increases the ambiguity

$$\overbrace{\text{katab} \qquad\qquad \text{katib} \qquad\qquad \text{kitab}}^{\text{ktb}}$$

katab



to write

katib

writer

kitab

book

Sometimes with completely different meanings:

$$\overbrace{\text{Eilom} \qquad\qquad \text{Ealam} \qquad\qquad \text{Ealim}}^{\text{Elm}}$$

Eilom
science

Ealam
flag

Ealim
to know

$\boxed{\text{1 Arabic token} = \text{proclitics} + \text{affixes} + \text{root} + \text{enclitics}}$

Example: وبحسناتهم (*wbHsnAthm* transliteration)
and by their virtues

enclitic	affix	stem	proclitics	
hm	At	**Hsn**	b	w
(their)	(s)	(virtue)	(by)	(and)

Language pair

Arabic: agglutination and segmentation

1 Arabic token = proclitics + affixes + root + enclitics

Example: وبحسناتهم (*wbHsnAthm* transliteration)
and by their virtues

enclitic	affix	stem	proclitics	
hm	At	**Hsn**	b	w
(their)	(s)	(virtue)	(by)	(and)

☞ Enclitics: pronouns and possessives

1 Arabic token = proclitics + affixes + root + enclitics

Example: وبحسناتهم (*wbHsnAthm* transliteration)
and by their virtues

enclitic	affix	stem	proclitics	
hm	At	**Hsn**	b	w
(their)	(s)	(virtue)	(by)	(and)

☞ Affixes: tense, genus and number marks

1 Arabic token = proclitics + affixes + root + enclitics

Example: وبحسناتهم (*wbHsnAthm* transliteration)
and by their virtues

enclitic	affix	stem	proclitics	
hm	At	**Hsn**	b	w
(their)	(s)	(virtue)	(by)	(and)

☞ Proclitics: prepositions, conjunctions and determiners

1. News domain

- Compilation of corpora supplied by LDC for the 2008 NIST Machine Translation Open Evaluation

1. News domain

- Compilation of corpora supplied by LDC for the 2008 NIST Machine Translation Open Evaluation

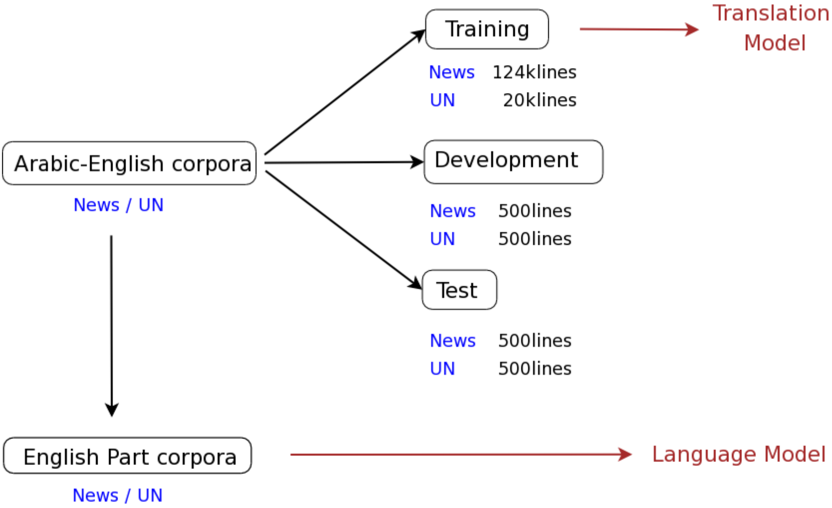
Corpus	Lines	Arabic tokens	English tokens
Arabic English Parallel News Part 1	61,000	2,179,289	2,273,021
Arabic News Translation Text Part 1	18,000	532,771	602,262
Arabic Treebank English Translation	23,800	660,821	739,695
eTIRR Arabic English News Text	4,000	97,882	98,655
Multiple-Translation Arabic	15,533	434,465	507,617
TIDES MT2004 Arabic evaluation data	1,329	40,667	47,324
Total:	123,662	3,945,895	4,262,740

Corpora
Parallel corpora

1. News domain
   - Compilation of corpora supplied by LDC for the 2008 NIST Machine Translation Open Evaluation
   - 123,662 lines, 3.9M Arabic tokens, 4.2M English tokens
2. United Nations transcriptions

Corpora
Parallel corpora

1. News domain
   - Compilation of corpora supplied by LDC for the 2008 NIST Machine Translation Open Evaluation
   - 123,662 lines, 3.9M Arabic tokens, 4.2M English tokens
2. United Nations transcriptions
   - Transcriptions from 1993 to 2002
   - Whole corpus: 3,686,372 lines
   - Training set: 20,000 lines, 805K Arabic tokens, 642K English tokens



Arabic-English corpora
News / UN
Training
News 124klines
UN 20klines
Translation Model
Development
News 500lines
UN 500lines
Test
News 500lines
UN 500lines
English Part corpora
News / UN
Language Model

Linguistic processing
Arabic

First step:
Buckwalter transliteration

- One to one correspondence between the Arabic glyphs and UTF-8
- Replace XML characters

<table>
<tr><td>ء</td><td>'</td><td>ذ</td><td>*</td><td>ل</td><td>l</td></tr>
<tr><td>آ</td><td>|</td><td>ر</td><td>r</td><td>م</td><td>m</td></tr>
<tr><td>أ</td><td>></td><td>ز</td><td>z</td><td>ن</td><td>n</td></tr>
<tr><td>ؤ</td><td>&</td><td>س</td><td>s</td><td>ه</td><td>h</td></tr>
<tr><td>إ</td><td><</td><td>ش</td><td>$</td><td>و</td><td>w</td></tr>
<tr><td>ئ</td><td>}</td><td>ص</td><td>S</td><td>ى</td><td>Y</td></tr>
<tr><td>ا</td><td>A</td><td>ض</td><td>D</td><td>ي</td><td>y</td></tr>
<tr><td>ب</td><td>b</td><td>ط</td><td>T</td><td>ـً</td><td>F</td></tr>
<tr><td>ة</td><td>p</td><td>ظ</td><td>Z</td><td>ـٌ</td><td>N</td></tr>
<tr><td>ت</td><td>t</td><td>ع</td><td>E</td><td>ـٍ</td><td>K</td></tr>
<tr><td>ث</td><td>v</td><td>غ</td><td>g</td><td>ـَ</td><td>a</td></tr>
<tr><td>ج</td><td>j</td><td>ـ</td><td>_</td><td>ـُ</td><td>u</td></tr>
<tr><td>ح</td><td>H</td><td>ف</td><td>f</td><td>ـِ</td><td>i</td></tr>
<tr><td>خ</td><td>x</td><td>ق</td><td>q</td><td>ـّ</td><td>~</td></tr>
<tr><td>د</td><td>d</td><td>ك</td><td>k</td><td>ـْ</td><td>o</td></tr>
</table>

Linguistic processing
Arabic

Original text:

وتأتي دول فرنسا وبريطانيا وايطاليا وألمانيا وأيرلندا وأسبانيا ولو كسمبورج في المقدمة
وبخلاف الشركات الأوروبية فقد وصل حجم رءوس الأموال المصدرة لل شركات العاملة في
مصر حتي ديسمبر ٢٠٠٠ الي ١٢٦ مليار ج نيه لعدد ١٠ آلاف شركة ا ستثماري ة
مؤسسة وفقا لقانون الاستثمار ..

Transliterated text:

```
wt>ty dwl frnsA wbryTAnyA wAyTAlyA w>lmAnyA w>yrlndA
w>sbAnyA wlwksmbwrj fy Almqdmp wbxlAf Al$rkAt Al>wrwbyp
fqd wSl Hjm r'ws Al>mwAl AlmSdrp ll$rkAt AlEAmlp fy mSr
Hty dysbmbr 2000 Aly 126 mlyAr jnyh lEdd 10 |lAf $rkp
AstvmAryp m&ssp wfqA lqAnwn AlAstvmAr ..
```

Linguistic processing
Arabic

Second step:
Tokenization

- Word segmentation in proclitics, stems+affixes, and enclitics
- Separates punctuation
- ASVMTools (Diab 2004) –except *Al-* determiner–

Example:

```
w t>ty dwl frnsA w bryTAnyA wAyTAlyA w >lmAnyA w >yrlndA w
>sbAnyA w lwksmbwrj fy Almqdmp w b xlAf Al$rkAt Al>wrwbyp
f qd wSl Hjm r'ws Al>mwAl AlmSdrp l Al$rkAt AlEAmlp fy
mSr Hty dysbmbr 2000 Aly 126 mlyAr jnyh l Edd 10 |lAf $rkp
AstvmAryp m&ssp wfqA l qAnwn AlAstvmAr
```

First + Second step $\Longrightarrow$ Linguistic tokens

Third step:
Annotate with Part-of-Speech and Chunk

- ASVMTools (Diab 2004)
- PoS: 24 tags (noun, adjective, verb...)
- Chunk: IOB tagging scheme (Inside-Outside-Beginning)
- Final text: *word*|*lemma*|*PoS*|*chunk*

Example:

w|w|CC|O tOty|tOty|VBP|B-VP dwl|dwl|NN|B-NP frnsA|frnsA|NNP|B-NP
w|w|CC|O bryTAnyA|bryTAnyA|NNP|B-NP wAyTAlyA|wAyTAlyA|JJ|I-NP w|w|CC|O
OlmAnyA|OlmAnyA|NNP|B-NP w|w|CC|O OyrlndA|OyrlndA|NNP|B-NP w|w|CC|O
OsbAnyA|OsbAnyA|NNP|B-NP w|w|CC|O lwksmbwrj|lwksmbwrj|NNP|B-NP
fy|fy|IN|B-PP Almqdmp|Almqdmp|NN|B-NP w|w|CC|B-PP b|b|IN|B-PP
xlAf|xlAf|NN|B-NP Al$rkAt|Al$rkAt|NNS|B-NP AlOwrwbyp|AlOwrwbyp|JJ|I-NP
f|f|CC|B-ADVP qd|qd|RP|B-PRT wSl|wSl|VBD|B-VP Hjm|Hjm|NN|B-NP
r'ws|r'ws|NN|B-NP AlOmwAl|AlOmwAl|NN|B-NP AlmSdrp|AlmSdrp|JJ|B-ADJP
l|l|IN|B-PP Al$rkAt|Al$rkAt|NNS|B-NP AlEAmlp|AlEAmlp|JJ|I-NP
fy|fy|IN|B-PP mSr|mSr|NNP|B-NP Hty|Hty|IN|B-PP dysbmbr|dysbmbr|NN|B-NP
2000|2000|CD|B-NP Aly|Aly|IN|B-PP 126|126|CD|B-NP mlyAr|mlyAr|NN|I-NP
jnyh|jnyh|NN|I-NP l|l|IN|B-PP Edd|Edd|NN|B-NP 10|10|CD|B-NP
LlAf|LlAf|NN|I-NP $rkp|$rkp|NN|I-NP AstvmAryp|AstvmAryp|JJ|I-NP
mWssp|mWssp|NN|B-NP wfqA|wfqA|NN|B-NP l|l|IN|B-PP qAnwn|qAnwn|NN|B-NP
AlAstvmAr|AlAstvmAr|NN|B-NP .|.|PUNC|O .|.|PUNC|O

First step:
Lowercase & Tokenization

Second step:
Part-of-Speech

- SVMTool (Giménez & Màrquez 2004)
- 36 tags (noun, adjective, verb...)

Third step:
Lemmatization

- Table (word,PoS) $\rightarrow$ lemma (185,201 entries)

Fourth step:
Chunking

- Yamcha (Kudo 2003)
- IOB tagging scheme (Inside-Outside-Beginning)
- Final text: *word*|*lemma*|*PoS*|*chunk*

Example:

```
france|france|NN|B-NP ,|,|,|I-NP britain|britain|NN|I-NP ,|,|,|O
italy|italy|RB|B-ADVP ,|,|,|O germany|germany|NN|B-NP ,|,|,|O
ireland|ireland|NN|B-NP ,|,|,|O spain|spain|NN|B-NP ,|,|,|O and|and|CC|O
luxembourg|luxembourg|NN|B-NP came|come|VBD|B-VP first.|first.|RB|B-ADVP
a|a|DT|B-NP part|part|NN|I-NP from|from|IN|B-PP the|the|DT|B-NP
european|european|JJ|I-NP companies|company|NNS|I-NP ,|,|,|O
the|the|DT|B-NP issued|issue|VBN|I-NP capital|capital|NN|I-NP
of|of|IN|B-PP companies|company|NNS|B-NP operating|operate|VBG|B-VP
in|in|IN|B-PP egypt|egypt|NN|B-NP reached|reach|VBN|B-VP
le126|le126|NN|B-NP billion|billion|CD|I-NP up|up|RP|B-ADVP
till|till|IN|B-PP december|december|NN|B-NP 2000.|2000.|CD|I-NP
such|such|JJ|I-NP capital|capital|NN|I-NP is|be|VBZ|B-VP of|of|IN|B-PP
10,000|10,000|CD|B-NP investment|investment|NN|I-NP
companies|company|NNS|I-NP set|set|VBN|B-VP up|up|RP|B-PRT
under|under|IN|B-PP the|the|DT|B-NP investment|investment|NN|I-NP
law|law|NN|I-NP .|.|.|O
```

- Language model
  - 5-gram Language Model, interpolated Kneser-Ney discounting
  - SRILM Toolkit (Stolcke 2002)
- Translation model
  - Alignments: GIZA++ Toolkit (Och & Ney 2003)
  - Translation tables: Moses package (Koehn et al. 2006)
    MLT package (Giménez & Màrquez)
- Decoder
  - Moses decoder (Koehn et al. 2006)
- Weights optimization
  - MERT (reference score BLEU)

- Language model
  - 5-gram Language Model, interpolated Kneser-Ney discounting
  - SRILM Toolkit (Stolcke 2002)
- ☞ Translation model
  - ☞ Alignments: GIZA++ Toolkit (Och & Ney 2003)
  - ☞ Translation tables: Moses package (Koehn et al. 2006)
MLT package (Giménez & Màrquez)
- Decoder
  - Moses decoder (Koehn et al. 2006)
- Weights optimization
  - MERT (reference score BLEU)

Statistical Machine Translation System
Building the system

- Language model
  - 5-gram Language Model, interpolated Kneser-Ney discounting
  - SRILM Toolkit (Stolcke 2002)
- Translation model
  - Alignments: GIZA++ Toolkit (Och & Ney 2003)
  - Translation tables: Moses package (Koehn et al. 2006)
MLT package (Giménez & Màrquez)
- ☞ Decoder
  - ☞ Moses decoder (Koehn et al. 2006)
- Weights optimization
  - MERT (reference score BLEU)

- Language model
  - 5-gram Language Model, interpolated Kneser-Ney discounting
  - SRILM Toolkit (Stolcke 2002)
- Translation model
  - Alignments: GIZA++ Toolkit (Och & Ney 2003)
  - Translation tables: Moses package (Koehn et al. 2006)
MLT package (Giménez & Màrquez)
- Decoder
  - Moses decoder (Koehn et al. 2006)

☞ Weights optimization

- ☞ MERT (reference score BLEU)

Results

Segmentation in the *News* compilation:

lAl$rkAt *vs.* l Al$rkAt *vs.* l Al $rkAt

	lines	tokens	toks/line
punct.	124,154	3,402,824	27.4
punct.+clitics	123,662	3,939,726	31.8
punct.+clitics+Al	123,498	4,718,933	38.2
English	123,662	4,262,740	34.5

Evaluation: BLEU (*n*-gram based metric)

For the three levels of segmentation:

	Arabic→English		English→Arabic	
	dev	test	dev	test
punct.	25.76	23.46	23.50	16.17
punct.+clitics	26.25	23.81	26.54	19.67
punct.+clitics+Al	25.28	23.21	32.46	26.68

Evaluation: BLEU (*n*-gram based metric)

For the three levels of segmentation:

	Arabic→English		English→Arabic	
	dev	test	dev	test
punct.	25.76	23.46	23.50	16.17
punct.+clitics	26.25	23.81	26.54	19.67
punct.+clitics+Al	25.28	23.21	32.46	26.68

☞ Best results when similar sentence lengths for both languages

Including linguistic information into a standard SMT

Including linguistic information into a standard SMT

Methods:

- Direct concatenation of the information
- Combination of two translation models
- One translation model with two factors

Including linguistic information into a standard SMT

Methods:

- Direct concatenation of the information
- Combination of two translation models
- One translation model with two factors

Best BLEU results: word&lemma

	Arabic→English		English→Arabic	
	dev	test	dev	test
w (baseline)	24.70	23.82	26.83	22.85
wl	24.74	24.28	26.95	23.34

Word Sense Disambiguation (WSD)

Identify the correct sense of a word given a sentence

Word Sense Disambiguation (WSD)

Identify the correct sense of a word given a sentence

$\Downarrow$

Different phrase translations $\equiv$ Different phrase senses

Word Sense Disambiguation (WSD)

Identify the correct sense of a word given a sentence

$\Downarrow$

Different phrase translations $\equiv$ Different phrase senses

$\Downarrow$

$\boxed{\text{Discriminative Phrase Translation (DPT)}}$

Discriminative phrase selection:

☞ Phrase selection is treated as a classification problem

- We use SVMs to solve the multiclass classification problem

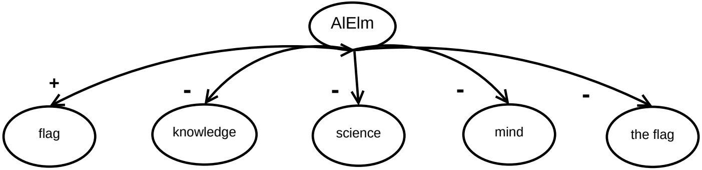
Discriminative phrase selection:

- Phrase selection is treated as a classification problem
- ☞ We use SVMs to solve the multiclass classification problem

Discriminative phrase selection:

- Phrase selection is treated as a classification problem
- We use SVMs to solve the multiclass classification problem

☞ Every possible translation is a class $\longrightarrow$ one-vs-all classification:



SMT system with discriminative phrase selection

The method

SVMs allow to use context and linguistic information

Features set for the SVMs include:

- Source phrase features
  - Word, PoS, coarse PoS and chunk *n*-grams
- Source sentence features
  - Word, PoS, coarse PoS, chunk *n*-grams and bag-of-words

For the previous example, AlElm:

Since the phrase is a word, phrase features are 1-grams:

Sentence :

w tAbE mr$d AlIxwAn “ In AlElm AlmTlwb fy dyn nA hw kl Elm nAfE tbqY l AlnAs vmrt h , swA' kAn ElmAF $rEyAF Ow ElmAF tjrybyAF .

Phrase features :

word *n*-grams	AlElm
PoS *n*-grams	NN
coarse PoS *n*-grams	N
chunk *n*-grams	B-NP

And the context of the whole sentence:

Sentence features:

word	(AlmTlwb)$_{1}$, (fy)$_{2}$, (dyn)$_{3}$, (nA)$_{4}$, (hw)$_{5}$, *n*-grams (" ln)$_{-2}$, (AllxwAn)$_{-3}$, (mr$d)$_{-4}$, (tAbE)$_{-5}$, (AlmTlwb fy)$_{1}$, (fy dyn)$_{2}$, (dyn nA)$_{3}$, (nA hw)$_{4}$, (ln AlmTlwb)$_{-1}$, (AllxwAn ")$_{-3}$, (mr$d AllxwAn)$_{-4}$, (tAbEmr$d)$_{-5}$ (AlmTlwb fy dyn)$_{1}$, (fy dyn nA)$_{2}$, (dyn nA hw)$_{3}$, (ln AlmTlwb fy)$_{-1}$, (" ln AlmTlwb)$_{-2}$, (AllxwAn " ln)$_{-3}$, (mr$d AllxwAn ")$_{-4}$, (tAbE mr$d AllxwAn)$_{-5}$
PoS	(JJ)$_{1}$, (IN)$_{2}$, (NN)$_{3}$, (PRP$)$_{4}$, (PRP)$_{5}$, *n*-grams (PUNC IN)$_{-2}$, (NN)$_{-3}$, (NN)$_{-4}$, (VBD)$_{-5}$ (JJ IN)$_{1}$, (IN NN)$_{2}$, (NN PRP$)$_{3}$, (PRP$ PRP)$_{4}$, (IN JJ)$_{-1}$, (NN PUNC)$_{-3}$, (NN NN)$_{-4}$, (VBD NN)$_{-5}$ (JJ IN NN)$_{1}$, (IN NN PRP$)$_{2}$, (NN PRP$ PRP)$_{3}$, (IN JJ IN)$_{-1}$, (PUNC IN JJ)$_{-2}$, (NN PUNC IN)$_{-3}$, (NN NN PUNC)$_{-4}$, (VBD NN NN)$_{-5}$,
coarse PoS *n*-grams	(J)$_{1}$, (I)$_{2}$, (N)$_{3}$, (P)$_{4}$, (P)$_{5}$, (P I)$_{-2}$, (N)$_{-3}$, (N)$_{-4}$, (V)$_{-5}$ (J I)$_{1}$, (I N)$_{2}$, (N P)$_{3}$, (P P)$_{4}$, (I J)$_{-1}$, (N P)$_{-3}$, (N N)$_{-4}$, (V N)$_{-5}$ (J I N)$_{1}$, (I N P)$_{2}$, (N P P)$_{3}$, (I J I)$_{-1}$, (P I J)$_{-2}$, (N P I)$_{-3}$, (N N P)$_{-4}$, (V N N)$_{-5}$
chunk *n*-grams	(I-NP)$_{1}$, (B-PP)$_{2}$, (B-NP)$_{3}$, (I-NP)$_{4}$, (B-NP)$_{5}$, (O B-SBAR)$_{-2}$, (B-NP)$_{-3}$, (B-NP)$_{-4}$, (B-VP)$_{-5}$ (I-NP B-PP)$_{1}$, (B-PP B-NP)$_{2}$, (B-NP I-NP)$_{3}$, (I-NP B-NP)$_{4}$, (B-SBAR I-NP)$_{-1}$, (B-NP O)$_{-3}$, (B-NP B-NP)$_{-4}$, (B-VP B-NP)$_{-5}$ (I-NP B-PP B-NP)$_{1}$, (B-PP B-NP I-NP)$_{2}$, (B-NP I-NP B-NP)$_{3}$, (B-SBAR I-NP B-PP)$_{-1}$, (O B-SBAR I-NP)$_{-2}$, (B-NP O B-SBAR)$_{-3}$, (B-NP B-NP O)$_{-4}$, (B-VP B-NP B-NP)$_{-5}$
bag-of-words	left: AllxwAn, mr$d, tAbE right: $rEyAF, AlmTlwb, AlnAs, Elm, ElmAF, dyn, kAn, kl, nAfE, swA', tbqY, tjrybyAF, vmrt

Estimation of the discriminative phrase translation model and integration into the SMT system:

- ☞ Training linear SVMs (SVM^{light}, Joachims 1999) for every translation of every phrase
- Convert SVM score into probability via a *softmax* function
- Include this probability in the translation model within a Log-linear model

Estimation of the discriminative phrase translation model and integration into the SMT system:

- Training linear SVMs (SVM^{light}, Joachims 1999) for every translation of every phrase
- ☞ Convert SVM score into probability via a *softmax* function
- Include this probability in the translation model within a Log-linear model

Estimation of the discriminative phrase translation model and integration into the SMT system:

- Training linear SVMs (SVM^{light}, Joachims 1999) for every translation of every phrase
- Convert SVM score into probability via a *softmax* function

☞ Include this probability in the translation model within a Log-linear model

Phrase translation task

Improvement in accuracy wrt. the most frequent translation, MFT

Occurrences	#	Acc.DPT (%)	Acc.MFT (%)
100-500	4,310	66.5	58.7
501-1,000	565	68.8	62.3
1,001-5,000	393	73.0	66.7
5,001-10,000	27	79.5	72.2
10,001-50,000	19	74.8	66.6
$>$ 50,000	7	80.7	76.2
Total:	5,321	67.3	59.8

Full translation task

Hyv$_{28}$ tm$_{22}$ AHrAq AlElm$_1$ AldnmArky .$_{1128}$

Translation table example:

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(e\|f)$
AlElm$_1$	flag	0.1986	0.3241
AlElm$_1$	the	0.0419	0.0207
AlElm$_1$	mind	0.0401	0.0620
AlElm$_1$	the flag	0.0397	0.0414
AlElm$_1$	flag during	0.0394	0.0138
AlElm$_1$	knowledge	0.0392	0.1103
AlElm$_1$	flag caused	0.0387	0.0138
AlElm$_1$	science	0.0377	0.1793
AlElm$_1$	education	0.0377	0.0138
AlElm$_1$	in mind	0.0371	0.0138

Full translation task

$\texttt{Hyv}_{28}$ $\texttt{tm}_{22}$ AHrAq $\texttt{AlElm}_{1}$ AldnmArky $._{1128}$

Translation table example:

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(e\|f)$
AlElm_1	flag	0.1986	0.3241
AlElm_1	the	0.0419	0.0207
AlElm_1	mind	0.0401	0.0620
AlElm_1	the flag	0.0397	0.0414
AlElm_1	flag during	0.0394	0.0138
AlElm_1	knowledge	0.0392	0.1103
AlElm_1	flag caused	0.0387	0.0138
AlElm_1	science	0.0377	0.1793
AlElm_1	education	0.0377	0.0138
AlElm_1	in mind	0.0371	0.0138

Three systems:

SMT	DPT	DPT$^+$
standard	replace MLE	add

SMT system with discriminative phrase selection

Evaluation

Three systems:

SMT	DPT	DPT$^+$
standard	replace MLE	add

Results:

	SMT	DPT	DPT$^+$
BLEU	23.88	23.87	23.96

☞ Non-significative improvements obtained

SMT system with discriminative phrase selection
Evaluation

Three systems:

SMT	DPT	DPT$^+$
standard	replace MLE	add

Results:

	SMT	DPT	DPT$^+$
BLEU	23.88	23.87	23.96

- Non-significative improvements obtained

☞ Coherent with other metrics, but in general, DPT better than DPT$^+$

Summary and conclusions

Conclusions

Summary

☞ First approach to the Arabic-to-English translation task

- Parallel corpora annotated with lemma, PoS and chunk
- Clitic segmentation improves the translation performance
- A direct inclusion of linguistic information (lemmas) slightly improves the results

Conclusions

Summary

- First approach to the Arabic-to-English translation task
- ☞ Parallel corpora annotated with lemma, PoS and chunk
- Clitic segmentation improves the translation performance
- A direct inclusion of linguistic information (lemmas) slightly improves the results

Conclusions

Summary

- First approach to the Arabic-to-English translation task
- Parallel corpora annotated with lemma, PoS and chunk
- ☞ Clitic segmentation improves the translation performance
- A direct inclusion of linguistic information (lemmas) slightly improves the results

Conclusions

Summary

- First approach to the Arabic-to-English translation task
- Parallel corpora annotated with lemma, PoS and chunk
- Clitic segmentation improves the translation performance

☞ A direct inclusion of linguistic information (lemmas) slightly improves the results

Conclusions

Summary

- ☞ We try to apply WSD techniques to the translation task. That allows to consider linguistic information and the context of the phrase

- Although we get an increment of 7.5% in accuracy in the phrase translation task, this is not reflected in full translation

- Necessity to improve the integration of DPT predictions into SMT system

Conclusions

Summary

- We try to apply WSD techniques to the translation task. That allows to consider linguistic information and the context of the phrase
- ☞ Although we get an increment of 7.5% in accuracy in the phrase translation task, this is not reflected in full translation
- Necessity to improve the integration of DPT predictions into SMT system

Conclusions

Summary

- We try to apply WSD techniques to the translation task. That allows to consider linguistic information and the context of the phrase
- Although we get an increment of 7.5% in accuracy in the phrase translation task, this is not reflected in full translation

☞ Necessity to improve the integration of DPT predictions into SMT system

Conclusions

Future work

☞ Complete DPT scores in the translation table when DPT prediction is not available

- Complement DPT predictions for both directions, $P_{\text{DPT}}(f|e)$ and $P_{\text{DPT}}(e|f)$, as done with MLE probabilities
- Deep into Arabic processing, especially in tokenization and lemmatization

Conclusions

Future work

- Complete DPT scores in the translation table when DPT prediction is not available

☞ Complement DPT predictions for both directions, $P_{\text{DPT}}(f|e)$ and $P_{\text{DPT}}(e|f)$, as done with MLE probabilities

- Deep into Arabic processing, especially in tokenization and lemmatization

Future work

- Complete DPT scores in the translation table when DPT prediction is not available
- Complement DPT predictions for both directions, $P_{\text{DPT}}(f|e)$ and $P_{\text{DPT}}(e|f)$, as done with MLE probabilities

☞ Deep into Arabic processing, especially in tokenization and lemmatization

Acknowledgments

OpenMT project of the *Ministerio de Educación y Ciencia*

Grup de Processament del Llenguatge Natural
Universitat Politècnica de Catalunya

...¡ GRÀCIES PEL SUPORT!

Jesús Giménez i Lluís Màrquez