Injected Linguistic Tags Approach to improve Phrase Based Statistical Machine Translation

Waleed Oransa, M.Sc.
PhD Candidate

Seminar in Laboratoire d'Informatique de l'Université du Maine
16 Sept 2011
Agenda

- Problem Statement
- The Methodology of the Study
- PBSMT Approach
- Online MT services review
- Arabic Linguistic issues in scope
- Injected Tags PBSMT Approach
- Experiments and Results
- Conclusion and Future work
Problem Statement

- English into Arabic translation generated by these MT systems has many Arabic linguistic issues. Some of these identified issues are:
  - Verb-Subject gender agreement
  - Verb-Subject number agreement
  - Noun-Adjective gender agreement
  - Noun-adjective number agreement
  - Counted Noun-Number agreement

- In short: issues related to **Gender and Number agreement**

- This research discusses these linguistic issues in the context of PBSMT with Arabic as example target language.
The Methodology of the Study

- This research is empirical, designed to achieve better machine translation.
- The Phrase Based Statistical Machine Translation (PBSMT) has been selected to be the base of this research due to its recent proven good translation quality over other approaches.
- The open source state-of-the-art PBSMT decoder “Moses” has been used.
- The automatic MT evaluation is based on the de-facto standard metric BLEU from NIST.
- The proposed approach has been also-evaluated by human compared to other online MT systems and rated according to the Arabic linguistic agreement rules in the scope.
The Methodology of the Study

- The used methodology consists of the following steps:
  1. Detect Mistakes
  2. Analyze These Mistakes
  3. Suggest an enhancement
  4. Implementation
  5. Test

Refinement

Keep

Yes

No

Test Result

MT Researcher

Human Assessor

Automatic Evaluation (e.g. BLEU)

Our System
Agenda

- Problem Statement
- The Methodology of the Study
- PBSMT Approach
- Online MT services review
- Arabic Linguistic issues in scope
- Injected Tags PBSMT Approach
- Experiments and Results
- Conclusion and Future work
PBSMT Approach

- Over all Phrase Pair Extraction Algorithm
  - Run a sentence aligner on a parallel bilingual corpus
  - Run word aligner (e.g., one based on IBM models) on each aligned sentence pair.
  - From each aligned sentence pair, extract all phrase pairs with no external links - see next slide

The scientist Einstein was born in 1879 A.D.

العالم اينشتئين ولد في سنة 1879 ميلادية
PBSMT Training Phase

**Input:** Training Corpus
Arabic/English Bi-Text

- **English Sentences**
- **Arabic Sentences**

**PBSMT System Training**

- **Translation Modeling Training**
  (Tool: Giza++ & Moses toolkit)
- **Language Modeling Training**
  (Tool: SRILM toolkit)

**Output:** Language Model and Translation Model

- **Translation Model Phrase Table**
  (English/Arabic)
- **Language Model**
  (Arabic)

PBSMT Normal Training Phase
PBSMT drawbacks

PBSMT has two drawbacks:

- Need a **huge corpus** to give good translation
- **Phrases are fragmenting the sentence** while in language like Arabic sentence should be homogenous (i.e. high dependencies between words and long distance inflection)
Agenda

- Problem Statement
- The Methodology of the Study
- PBSMT Approach
- Online MT services review
- Arabic Linguistic issues in scope
- Injected Tags PBSMT Approach
- Experiments and Results
- Conclusion and Future work
<table>
<thead>
<tr>
<th></th>
<th>Sentence/Translation Service</th>
<th>Arabic Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Fifteen girls</td>
<td>خمس عشرة فتاة</td>
</tr>
<tr>
<td>1</td>
<td>Google Translate (GO)</td>
<td>خمسة عشر فتيات</td>
</tr>
<tr>
<td>2</td>
<td>MS-Bing Translator (MS)</td>
<td>خمسة عشر الفتيات</td>
</tr>
<tr>
<td>3</td>
<td>Systran translator (SY)</td>
<td>خمسة عشر بنات</td>
</tr>
<tr>
<td>4</td>
<td>Sakhr Trjem (SK)</td>
<td>خمسة عشر بنات</td>
</tr>
<tr>
<td>B</td>
<td>The two girls said &quot;we are good&quot;</td>
<td>قالت الفتاتان &quot;نحن جيدات&quot;</td>
</tr>
<tr>
<td>5</td>
<td>GO</td>
<td>الفتياتن وقال &quot;نحن جاهزون&quot;</td>
</tr>
<tr>
<td>6</td>
<td>MS</td>
<td>إن فتاتين &quot;نحن جيدة&quot;</td>
</tr>
<tr>
<td>7</td>
<td>SY</td>
<td>قال الاتنان بنات &quot;نحن جيد&quot;</td>
</tr>
<tr>
<td>8</td>
<td>SK</td>
<td>البناتن قالتا أثنا جيدين</td>
</tr>
</tbody>
</table>
Agenda

- Problem Statement
- The Methodology of the Study
- PBSMT Approach
- Online MT services review
- Arabic Linguistic issues in scope
- Injected Tags PBSMT Approach
- Experiments and Results
- Conclusion and Future work
Arabic Linguistic issues in scope

- Arabic Gender & numbers (singular, dual, plural):

<table>
<thead>
<tr>
<th>Gender</th>
<th>S/D/P</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine</td>
<td>Singular</td>
<td>man رجل</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>(two men) رجلان</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>men رجال</td>
</tr>
<tr>
<td>Feminine</td>
<td>Singular</td>
<td>girl فتاة</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>(two girls) فتيات</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>girls فتيات</td>
</tr>
<tr>
<td>Proper Name</td>
<td>Singular</td>
<td>Waleed وليد</td>
</tr>
<tr>
<td>(male)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proper Name</td>
<td>Singular</td>
<td>Rihan ريهان</td>
</tr>
<tr>
<td>(female)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Arabic Linguistic issues in scope

- Noun and cardinal numerals
  1. Numerals 1 and 2: Agree with the noun's gender.
  2. Numerals 3 -10: Reverse with the noun's gender.
  3. Numerals 11-12: Agree with the noun's gender.
  5. Numerals 20-99: first part (apply rule 1 and 2), second part "عشرون" is always masculine regardless of the counted noun's gender.
Arabic Linguistic issues in scope

- Noun and cardinal numerals (noun inflection)
  1. Numeral 1: singular noun.

<table>
<thead>
<tr>
<th>English</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>One man won</td>
<td>فاز رجل واحد</td>
</tr>
<tr>
<td>One girl won</td>
<td>فازت فتاة واحدة</td>
</tr>
<tr>
<td>The man has three sons</td>
<td>للرجل ثلاثة أولاد</td>
</tr>
<tr>
<td>Fourteen rooms</td>
<td>أربع عشرة غرفة</td>
</tr>
</tbody>
</table>
Arabic Linguistic issues in scope

- Verb inflection based on gender and number
  - An example of the verb inflection is adding the suffix "ته " to the past tense of the verb for singular feminine subject.

<table>
<thead>
<tr>
<th>English</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man went to home</td>
<td>الرجل ذهب إلى المنزل</td>
</tr>
<tr>
<td>The girl went to home</td>
<td>الفتاة ذهبت إلى المنزل</td>
</tr>
<tr>
<td>The two men went to home</td>
<td>الرجال ذهبا إلى المنزل</td>
</tr>
<tr>
<td>The two girls went to home</td>
<td>الفتيات ذهبتا إلى المنزل</td>
</tr>
<tr>
<td>The men went to home</td>
<td>الرجال ذهبوا إلى المنزل</td>
</tr>
<tr>
<td>The girls went to home</td>
<td>الفتيات ذهبن إلى المنزل</td>
</tr>
</tbody>
</table>
### Arabic Linguistic issues in scope

- Inflection of gender and number on other nouns: They have inflection on other nouns like adjectives, apposition, adverbs.

<table>
<thead>
<tr>
<th>English</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>The tall boy</td>
<td>الولد الطويل</td>
</tr>
<tr>
<td>The tall girl</td>
<td>الفتاة الطويلة</td>
</tr>
<tr>
<td>The two tall men</td>
<td>الرجال الطويلان</td>
</tr>
<tr>
<td>The two tall girls</td>
<td>الفتيات الطويلتان</td>
</tr>
<tr>
<td>The tall men</td>
<td>الرجال الطوال</td>
</tr>
<tr>
<td>The tall girls</td>
<td>الفتيات الطويلات</td>
</tr>
</tbody>
</table>
Agenda

- Problem Statement
- The Methodology of the Study
- PBSMT Approach
- Online MT services review
- Arabic Linguistic issues in scope
- Injected Tags PBSMT Approach
- Experiments and Results
- Conclusion and Future work
The proposed approach focuses on strengthen the SMT capabilities to encode more linguistic information.

This approach utilizes nouns' similarity.

Similarity here means similar in gender, number (and may be other linguistic features in future work).

The desired noun's similarity should be derived from the target language morphology and features with some assistance from source language.

<table>
<thead>
<tr>
<th>The man went to home</th>
<th>الرجل ذهب للمنزل</th>
</tr>
</thead>
<tbody>
<tr>
<td>The boy went to home</td>
<td>?</td>
</tr>
</tbody>
</table>
The IT's approach consists of two-phases:

- Training phase: includes preprocessing and PBSMT training.
- Translation phase: includes preprocessing, PBSMT decoding, post processing tags decoding and translation.
Injected Tags Training Phase

Input: Training Corpus Arabic/English Bi-Text

English Sentences

Text Preprocessing

Dictionary

Arabic Sentences

English Sentences with ITs

Arabic Sentences with ITs

Translation Modeling Training (Tool: Giza++ & Moses toolkit)

Language Modeling Training (Tool: SRILM toolkit)

PBSMT System Training

Translation Model (English/Arabic)

Language Model (Arabic)

Output: Language Model and Translation Model

16 Sept 2011 - Page 21
Injected Tags PBSMT Approach

- Injected Tags (ITs) represent the linguistic features associated with these features:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number</th>
<th>Injected Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine Noun</td>
<td>Singular</td>
<td>NM</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>NMT</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>NMM</td>
</tr>
<tr>
<td>Feminine Noun</td>
<td>Singular</td>
<td>NF</td>
</tr>
<tr>
<td></td>
<td>Dual</td>
<td>NFT</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>NFM</td>
</tr>
<tr>
<td>Proper Name (male)</td>
<td>Singular</td>
<td>NM</td>
</tr>
<tr>
<td>Proper Name (Female)</td>
<td>Singular</td>
<td>NF</td>
</tr>
</tbody>
</table>
Training phase

1) Inject Tags (ITs) in preprocessing Phase.

<table>
<thead>
<tr>
<th>Description</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row English/Arabic parallel text</td>
<td>the women who drive the car</td>
</tr>
<tr>
<td></td>
<td>النساء الآتية يقودن السيارة</td>
</tr>
<tr>
<td>Preprocessing output</td>
<td>the NFM who drive the NF</td>
</tr>
<tr>
<td></td>
<td>NF الآتية يقودن ال NFM</td>
</tr>
</tbody>
</table>
### Injected Tags PBSMT Approach

#### Training phase

2) ITs Normalization (derived Tags)

<table>
<thead>
<tr>
<th>Rule</th>
<th>Example</th>
</tr>
</thead>
</table>
| X \(\ldots\) X \(\rightarrow\) X  
X بن X \(\rightarrow\) X               | NF NF \(\rightarrow\) NF  
NM بن NM \(\rightarrow\) NM           |
| NF \(\ldots\) X \(\rightarrow\) NF  | NF NM \(\rightarrow\) NF  |
| NM and X \(\rightarrow\) NMT       | NM and NF \(\rightarrow\) NMT |
| NM و X \(\rightarrow\) NMT          |                          |
| X and X \(\rightarrow\) XT        | NF and NF \(\rightarrow\) NFT  
X و X \(\rightarrow\) XT               | NM و NM \(\rightarrow\) NMT   |
| NM and X \(\ldots\) \(\rightarrow\) NMM  
NM و X \(\ldots\) \(\rightarrow\) NMM | NM and NF \(\rightarrow\) NMM |
| X and X \(\ldots\) \(\rightarrow\) XMM  
X و X \(\ldots\) \(\rightarrow\) XMM | NF and NF \(\rightarrow\) NFM  
NM و NM \(\rightarrow\) NMM           |
| X or X \(\ldots\) \(\rightarrow\) X  
X و X \(\ldots\) \(\rightarrow\) X | NF or NF \(\rightarrow\) NF  
NM و NM \(\rightarrow\) NM           |
### Injected Tags PBSMT Approach

**Training phase**

3) PBSMT training

<table>
<thead>
<tr>
<th>Description</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row English/Arabic parallel text</td>
<td>Sarah Foster can drive the car</td>
</tr>
<tr>
<td></td>
<td>سارة فوستر تستطيع قيادة السيارة</td>
</tr>
<tr>
<td>Preprocessing output</td>
<td>NF NM can drive the NF</td>
</tr>
<tr>
<td></td>
<td>تستطيع قيادة NF الفنون</td>
</tr>
<tr>
<td>Normalized</td>
<td>NF can drive the NF</td>
</tr>
<tr>
<td></td>
<td>تستطيع قيادة الفنون</td>
</tr>
</tbody>
</table>
ITs Training Phase Example

Input: Training Corpus Arabic/English Bi-Text

Text Preprocessing

PBSMT System Training

Translation Modeling Training (Tool: Giza++ & Moses toolkit)

Language Modeling Training (Tool: SRILM toolkit)

Output: Language Model and Translation Model

Sara Foster can drive the car

Dictionary Arabic-English

NF can drive the NF

Language Model (Arabic)

Translation Model (English/Arabic)
Injected Tags PBSMT Approach

Translation phase

1) Preprocessing for source text.
2) PBSTM decoding.
3) Post Processing (include de-normalization).
4) Translation of existing Injected Tags (ITs).
ITs Translation Phase

Source Text

English Sentence

Text Preprocessing

Dictionary Arabic-English

English sentence with ITs

PBSMT System (Decoder)

Arabic Sentence with ITs

Text Post Processing Replace ITs

Arabic sentence

Target Text

ITs Translation Table

IT – En- Ar
IT – En- Ar
IT – En- Ar
ITs Translation Phase

Source Text

John and Mike bought a new car

Text Preprocessing

NMT bought a new NF

PBSMT System (Decoder)

Text Post Processing

Replace ITs

Target Text

16 Sept 2011 - Page 29
Agenda

- Problem Statement
- The Methodology of the Study
- PBSMT Approach
- Online MT services review
- Arabic Linguistic issues in scope
- Injected Tags PBSMT Approach
- Experiments and Results
- Conclusion and Future work
Experiments

- We developed a small “noun/proper noun” dictionary to be used in the experiments.
- Evaluate the new system that partially using ITs approach compared to the baseline system with no ITs and show the impact on the translation quality.
- Used two corpora: 4K, 37K sentences
Experiment 1

- Manually develop corpus and noun dictionary for a few hundred sentences
- Three techniques: ITC1, ITC2, ITC3
- ITC1: keeps the replaced Arabic/English aligned nouns inside the corpus but in separate line
- ITUN & ITC2: Removes the replaced Arabic/English aligned nouns from the parallel text
Experiment 1

- ITC3: Same as ITC2 but in addition to keep the original sentences in the corpus without any processing
- Train the PBSMT system with the inject parallel sentences.
- Evaluate output in terms of BLEU
ITUN & ITC2 Technique

Input: Training Corpus
Arabic/English Bi-Text

Text Preprocessing

Dictionary
Arabic-English

Translation Modeling Training
(Tool: Giza++ & Moses toolkit)

Language Modeling Training
(Tool: SRILM toolkit)

PBSMT System Training

Translation Model
(English/Arabic)

Language Model
(Arabic)

Output: Language Model and Translation Model

Aliaa Nabil can drive the car
علياء نبيل تستطيع قيادة السيارة

NF can drive the NF
NF تستطيع قيادة NF

NF can drive the NF
NF can drive the NF
ITC1 Technique

Input: Training Corpus
Arabic/English Bi-Text

Text Preprocessing

NF can drive the NF

Dictionary
Arabic-English

Aliaa Nabil can drive the car

Translation Modeling Training
(Tool: Giza++ & Moses toolkit)

Language Modeling Training
(Tool: SRILM toolkit)

PBSMT System Training

Translation Model (English/Arabic)

Language Model (Arabic)

Output: Language Model and Translation Model

Aliaa Nabil can drive the car

NF can drive the NF

Aliaa Nabil can drive the car

NF can drive the NF

16 Sept 2011 - Page 35
ITC3 Technique

Input: Training Corpus Arabic/English Bi-Text

Text Preprocessing

Translation Modeling Training (Tool: Giza++ & Moses toolkit)
Language Modeling Training (Tool: SRILM toolkit)

PBSMT System Training

Output: Language Model and Translation Model
## Experiment 1 Results

**Evaluation results using BLEU score:**

<table>
<thead>
<tr>
<th>System</th>
<th>Original Corpus size</th>
<th>BLEU score</th>
<th>Number of Injected tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (BL)</td>
<td>4k</td>
<td>0.0993</td>
<td>-</td>
</tr>
<tr>
<td>Injected Tags&lt;sub&gt;1&lt;/sub&gt; Corpus (ITC1)</td>
<td>4k</td>
<td>0.1884</td>
<td>1,156</td>
</tr>
<tr>
<td>Injected Tags&lt;sub&gt;2&lt;/sub&gt; Corpus (ITC2)</td>
<td>4k</td>
<td>0.1902</td>
<td>1,156</td>
</tr>
<tr>
<td>Injected Tags&lt;sub&gt;3&lt;/sub&gt; Corpus (ITC3)</td>
<td>4k</td>
<td>0.2013</td>
<td>1,156</td>
</tr>
<tr>
<td>Baseline (BL-UN)</td>
<td>37k</td>
<td>0.6171</td>
<td>-</td>
</tr>
<tr>
<td>Injected Tags Corpus (ITUN)</td>
<td>37k</td>
<td><strong>0.6979</strong></td>
<td><strong>13,127</strong> (13% increase)</td>
</tr>
</tbody>
</table>
Experiment 1 Results

For 4k corpus

BLEU Score

Baseline: 0.0993
ITC1: 0.1884
ITC2: 0.1902
ITC3: 0.2013
Experiment 1 Results

For 37k corpus (UN open corpus)

BLEU Score

Baseline: 0.6171
ITUN: 0.6979

13% Improvement
Experiment 2

- Manually develop corpus and noun dictionary for a few hundred sentences
- Develop a test data that doesn’t contain any sentence from the training data
- Train the PBSMT system with the prepared training data
- Translate the test data using:
  - The trained ITs PBSMT System.
  - The online MT services.
- Evaluate output of all systems by **human assessor**
## Experiment 2 Results

**Gender & number agreement human evaluation results**

<table>
<thead>
<tr>
<th>1- The rich man drives the car</th>
<th>2- Four girls were walking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GO</strong></td>
<td>رجل غني محركات السيارة</td>
</tr>
<tr>
<td><strong>MS</strong></td>
<td>الرجل الغني محركات السيارة</td>
</tr>
<tr>
<td><strong>FT</strong></td>
<td>للأغنيء تقود السيارة في</td>
</tr>
<tr>
<td><strong>SY</strong></td>
<td>يقود الرجل غني السيارة</td>
</tr>
<tr>
<td><strong>SK</strong></td>
<td>الرجل الغني يقود السيارة</td>
</tr>
<tr>
<td><strong>IT</strong></td>
<td>الرجل الغني يقود السيارة</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3- This woman read the book</th>
<th>4- Three boys were happy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GO</strong></td>
<td>هذه المرأة قراءة الكتاب</td>
</tr>
<tr>
<td><strong>MS</strong></td>
<td>هذه امرأة قراءة الكتاب</td>
</tr>
<tr>
<td><strong>FT</strong></td>
<td>هذه امرأة تقرأ الكتاب</td>
</tr>
<tr>
<td><strong>SY</strong></td>
<td>هذا إمرأة قرأ الكتاب</td>
</tr>
<tr>
<td><strong>SK</strong></td>
<td>قرأت هذه المرأة الكتاب</td>
</tr>
<tr>
<td><strong>IT</strong></td>
<td>هذه المرأة قرأت الكتاب</td>
</tr>
</tbody>
</table>
## Experiment 2 Results

- **Gender & number agreement human evaluation results**

<table>
<thead>
<tr>
<th></th>
<th>5- Fifteen boys said good man</th>
<th>6- The two boys were playing</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO</td>
<td>الأولاد خمسة عشر قال رجل جيد</td>
<td>اثنين من الفتيان كانوا يلعبون</td>
</tr>
<tr>
<td>MS</td>
<td>وقال الفتيان خمسة عشر رجل جيد</td>
<td>كانوا يلعبون صبيان</td>
</tr>
<tr>
<td>FT</td>
<td>وقال أحد الصبية جيدة Fifteen</td>
<td>وكان اثنان الصبية كانوا يلعبون</td>
</tr>
<tr>
<td>SY</td>
<td>خمسة عشر فتيان يقول جيد رجل</td>
<td>الأثنان فتيان كان لعب</td>
</tr>
<tr>
<td>SK</td>
<td>قال خمسة عشر ولد الرجل الجيد</td>
<td>الوُلَدَان كانا يلعبان</td>
</tr>
<tr>
<td>IT</td>
<td>خمسة عشر فتيان قالوا رجل جيد</td>
<td>الوُلَدَان كانا يلعبان</td>
</tr>
</tbody>
</table>

16 Sept 2011 - Page 42
Experiment 2 Results

- Gender & number agreement human evaluation results

![Bar chart showing results for Google Translate (14%), Sakhr Tarjem (57%), and Injected Tags (85%)](chart.png)
Agenda

- Problem Statement
- The Methodology of the Study
- PBSMT Approach
- Online MT services review
- Arabic Linguistic issues in scope
- Injected Tags PBSMT Approach
- Experiments and Results
- Conclusion and Future work
Conclusion

- Injected Tags Approach is presented which utilize the noun similarity and provides Arabic linguistic information into PBSMT systems.
- The approach can be used with any other language pair to resolve the PBSMT limitations related to linguistic information.
- This approach also provides a robust technique to expand the available corpus with custom dictionary that enriches corpora of PBSMT and so it decreases the out-of-vocabulary rate.
Conclusion

- A machine translation system uses this approach has been developed and evaluated using BLEU automatic evaluation metric and by using human evaluation compared to other online MT services.
- The ITs approach improves the quality of the translation by respecting Arabic noun's gender-number agreement with other sentence parts.
- The approach gives a relative increase of more than 13% of the BLEU score compared to the baseline system.
Future Work

- More enhancements could be done in order to achieve better results through:
  - Addressing the current discovered issues
  - Extending the current injected tags to address more linguistic information
  - Enhancing the architecture to improve performance and scalability
This presentation illustrates the original work published as follows:


IEEEXplore Link:
http://ieeexplore.ieee.org/search/srchabstract.jsp?tp=&arnumber=5451681
Thank you شكرآً