Using Verb Paraphrases in Arabic-to-English Example-Based Translation

Kfir Bar
Nachum Dershowitz
Tel Aviv University
MTML 2011
EBMT – Example Based Machine Translation

Paraphrases

Matching

Transfer

Recombination

Source language text

Bi-lingual corpus

Target language text
Our EBMT System

- Non-structured: translation examples are stored with only some morph-syntactic information
- Uses a parallel corpus provided by LDC
- So far, only matching and transfer. Real recombination left for future work
Corpus

- Uses sentence-aligned parallel corpus
- Translation examples were
  - morphologically analyzed using Buckwalter
  - part-of-speech tagged using AMIRA
- Word alignment by GIZA++
Matching

- Corpus is searched for input fragments
- Matching is word-by-word at several levels.
  Total score is calculated by combining level scores

- Exact match
  - Single-word Paraphrase match
  - Stem match
  - Lemma match
  - Morphological-feature match

- Fragment score is created from word scores
Input sentence:
مذكرة من رئيس مجلس الأمن
(A memorandum by the president of the Security Council)

Corpus example:
... ويعين مجلس الوزراء أعضاء اللجنة ويجري...
Inspired by:

*Extracting Paraphrases from a Parallel Corpus, Regina Barzilay and Kathleen R. McKeown*

Multiple **English translations of same source**

- Extracted multi-word paraphrases
- Co-trained two classifiers –
  one for best paraphrase contexts
  one for best paraphrase patterns
Two verbs expressing the same meaning

In at least one context the two are understood in the same way

The radio broadcast these songs
Siniora's office and Olmert’s administration deny a story about a meeting in Sharm al-Sheikh.

Siniora's office denies a story about a meeting with Israeli officials.

**Verb:** *nfy* (*ynfyAn, ynfy*)

**Context:**
Left-1 (NN, NNP) Right-1 (NN₁, IN₂)
Left-2 (NN, NNP) Right-2 (NN₁, IN₂)
Paraphrase Extraction

Arabic Verbs

root + inflectional class (9)
$.r.k xaAxaxa = $aAraka

Inflections
- Gender (m, f)
- Person (1, 2, 3)
- Number (singular, dual, plural)

Verb
- Perfect
  - $aArak
- Imperfect
  - $aArik

stem $aArak
lemma $aArak_1
Arabic has some features that may affect paraphrasing:

- Using different particles indicating the object may change the meaning of the verb
  
  \( qDY = \) to judge
  
  \( qDY \ EIY = \) to kill

- In some cases, the verb can be replaced with its verbal noun
  
  \( zAr = \) to visit = \( qAm \) \( bzyArp \)
  
  \( wqE = \) signed = \( tm \) \( twqyE \)
Sometime one cannot identify the voice, since diacritics are omitted

*kutiba* (written) and *kataba* (to write)

Class 7 is the passive version of class 1 and class 5 is the passive version of class 2

*qutiEa* (cut - passive) = *linqataE*
Paraphrase Extraction

- **Using parallel corpus:**
  
  *Improved statistical machine translation using paraphrases, Chris Callison-Burch, Philipp Koehn and Miles Osborne*

- **Using large monolingual corpus:**

  *Discovery of Inference Rules for Question Answering, Dekang Lin and Patrick Pantel*

- **Using corpus of comparable documents:**

  *Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment, Regina Barzilay and Lillian Lee*
First attempt in Arabic – using comparable corpora

Comparable corpora extracted from the Arabic Gigaword. We automatically paired documents that were published on the same day and share title word lemmas
Paraphrase Extraction

Corpus of Comparable Documents

\[
similar(A, B) / \text{length(title}(A)) = 0.2
\]
Paraphrase Extraction

Extracting contexts of positive examples and negative examples, and learning the best contexts for both classes.

Step 1

- We consider every verb-pair as a candidate.
- We choose all candidates whose verbs share a lemma, as positive examples.
- Negative examples are candidates with unrelated verbs.
- The rest are potential candidates.
We created a list of related verbs:

Buckwalter’s **stem list** comes with English glosses.

- We relate verbs that share either similar or synonymous English translations.
- Synonyms were found using the English WordNet.
Identifying the best contexts using the strength and frequency of each context

\[ \text{strength}(C) = \frac{X}{N} \]

\( X \) is the number of times the context \( C \) appears in a positive / negative example

\( N \) is the total number of occurrences of \( C \)
For each class (negative / positive) we choose the \textbf{\textit{k most frequent}} contexts with \textit{strength} higher than a predefined threshold \( k=20, \ \text{threshold}=0.95 \).

We consider all occurrences of the best contexts among all potential candidates.

\textit{Paraphrases} are those surrounded by a \textbf{positive context} but not surrounded by a \textbf{negative one}.
We used 5500 document pairs, extracted from the Arabic Gigaword. Total number of words is ~3M.

In this experiment we use a **window of size 2**
2 experts evaluated the results. For each candidate pair, every expert was requested to decide:

correct – instances can be exchanged in some contexts
incorrect – instances are not exchangeable

<table>
<thead>
<tr>
<th># of Unique Candidates</th>
<th># of Unique Paraphrases</th>
<th>Expert 1: # of correct Paraphrases</th>
<th>Expert 2: # of correct Paraphrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,101</td>
<td>139</td>
<td>120 (86% precision)</td>
<td>103 (74% precision)</td>
</tr>
</tbody>
</table>
Paraphrase Extraction

Examples

Paraphrases

AEtql  ↔ Owqf
bv~  ↔ nq~l
Istqbl  ↔ IltqY

Best Positive Contexts

L NN₀ ↔ NN₀  R IN, NN ↔ IN
L NN, WP₀ ↔ WP₁  R NN₀ ↔ NN₀
Paraphrases in Translation

- **Exact match**
- **Single-word Paraphrase match**
- **Stem match**
- **Lemma match**
- **Morphological-feature match**

- Testing on 586 sentences (MT-EVAL 09)
- Size of parallel corpus: ~ 1.2 Million Arabic tokens
**Paraphrases in Translation**

| Para-fragments in final translations | 63 |
| Translated Para-fragments | 97 |
| Uncovered Para-fragments | 57 |
| Para-fragments | 193 |

*Para-fragment* = a fragment based on a paraphrase-match

**Evaluation**

**BLEU**

<table>
<thead>
<tr>
<th></th>
<th>15.06</th>
<th>15.10</th>
</tr>
</thead>
</table>

**Translation**

- Good: 42
- Bad: 21
- Total: 162 correct
Conclusions

- Single-word verb paraphrases can help in matching Arabic input text with translation examples.
- We believe that involving more paraphrases in translation will improve results, especially when faced with a small parallel corpus.
- The paraphrasing classifier achieves a relatively high precision but maybe lower recall. Other techniques should be tested.
Future Work

- Add more features to the paraphrase extraction process (e.g. “sibling” content words)
- Consider using a dependency syntax parser in the paraphrase extraction process
- Find multi-word Arabic paraphrases and use them in translation
Thank you