Using Synonyms for Arabic-to-English Example-Based Translation

Kfir Bar
Nachum Dershowitz
Tel Aviv University

AMTA 2010
EBMT – Example Based Machine Translation

- **Matching**
  - Synonyms

- **Transfer**

- **Recombination**
  - Target language text

Source language text

Bi-lingual corpus
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.
Translation examples were morphologically analyzed using the Buckwalter morphological analyzer, and then part-of-speech tagged using AMIRA (Diab et al., 2004).

Word alignment in each translation example is done by GIZA++. Unaligned words were aligned using a lexicon enriched with WordNet synonyms.

Corpus

- Uses sentence-aligned parallel corpus (by LDC).
- Translation examples were morphologically analyzed using the Buckwalter morphological analyzer, and then part-of-speech tagged using AMIRA (Diab et al., 2004).
- Word alignment in each translation example is done by GIZA++. Unaligned words were aligned using a lexicon enriched with WordNet synonyms.
Corpus is searched for input fragments.

Matching is word-by-word at several levels. Total score is calculated by combining level scores.

Fragment score is created from word scores.
There are several works on automatic extraction of synonyms and semantically similar expressions:

- *Translations as Semantic Mirrors: From Parallel Corpus to WordNet*, Dyvik Helge. 2004
- *Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity*, Lonneke van der Plas and Jörg Tiedemann. 2006

Our current attempt uses Buckwalter lexicon and English WordNet (EWN) for finding Arabic noun synonyms.
Every noun stem in the Buckwalter list was compared to all other stems

We ask EWN for all (noun) synsets of every English translation of a stem.

A synset containing two or more Buckwalter translations is a possible sense for the stem.
We also considered the hypernym relation.
We define **five** levels of synonymy between **stems**:

1. 2 or more translations in common
2. 1 or more senses in common
3. Same unique translation
4. 1 translation each and they’re synonyms
5. 1 common translation
Thesaurus Extraction

Example:

AEAdp

Level 2

repetition

return

recurrence

krwr
### Thesaurus Extraction

**Example:**

<table>
<thead>
<tr>
<th>Synonyms</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>n$yj / dmE (“crying”)</td>
<td>4</td>
</tr>
<tr>
<td>sTH / sqf (“ceiling”)</td>
<td>5</td>
</tr>
<tr>
<td>zlEwm / Hlqwm (“throat”)</td>
<td>1</td>
</tr>
<tr>
<td>njdp / AEAnp (“help;support”)</td>
<td>2</td>
</tr>
<tr>
<td>AbtdA' / ftH (“beginning”)</td>
<td>5</td>
</tr>
<tr>
<td>AxtrAE / AbtkAr (“invention”)</td>
<td>3</td>
</tr>
</tbody>
</table>
The resultant thesaurus contains:

22,621 nouns

1. 20,512 relations
2. 1,479 relations
3. 17,166 relations
4. 38,754 relations
5. 137,240 relations
Since words in the input sentence / corpus are not given with their senses it is difficult to match on synonyms.

- Use word-sense-disambiguation tool for Arabic
- Use local context to find if two words may be synonyms

We classify each input sentence by topic, as well as all the corpus translation examples. We consider synonyms only if the topic-sets of both parts intersect.
Classification

We trained a simple classifier on English Reuters corpus.

We used SVM on stems, removing stop words. **Accuracy: 94% for Reuters test-set (1219 documents).**

Used classifier on English half of all translation examples in our corpus.

The Arabic part of those examples was used as a training-set for another classifier for the same topic list for Arabic (stems, ignoring stop words).
## Results

<table>
<thead>
<tr>
<th></th>
<th>Small Corpus</th>
<th>Large Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29,992 translation examples</td>
<td>58,115 translation examples</td>
</tr>
<tr>
<td>w/ classification</td>
<td>w/o classification</td>
<td>w/ classification</td>
</tr>
<tr>
<td>Level 1</td>
<td>0.1186</td>
<td>0.1176</td>
</tr>
<tr>
<td>Levels 1 – 2</td>
<td>0.1176</td>
<td>0.1173</td>
</tr>
<tr>
<td>Levels 1 – 3</td>
<td>0.1186</td>
<td>0.1176</td>
</tr>
<tr>
<td>Levels 1 – 4</td>
<td>0.1187</td>
<td>0.1179</td>
</tr>
<tr>
<td>Levels 1 – 5</td>
<td><strong>0.1192 (+9%)</strong></td>
<td>0.1177</td>
</tr>
<tr>
<td>No synonym</td>
<td></td>
<td>0.1084</td>
</tr>
</tbody>
</table>

Testing on 586 sentences (MT-EVAL 09)
### Results

Uncovered N-grams in the small corpus

<table>
<thead>
<tr>
<th>N-grams</th>
<th>w/ synonyms</th>
<th>w/o synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigrams</td>
<td>733</td>
<td>738</td>
</tr>
<tr>
<td>Bigrams</td>
<td>7612 (+3.2%)</td>
<td>7864</td>
</tr>
<tr>
<td>Trigrams</td>
<td>11554</td>
<td>11632</td>
</tr>
<tr>
<td>4-grams</td>
<td>11224</td>
<td>11243</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

- Synonyms benefit from being matched carefully by considering the context in which they appear.
- Using synonyms on a large corpus did not result in an improvement of the final results, as it did for a smaller corpus.
- Improving alignment and smoothing out the final English translation is under development.
- Beginning to investigate the possibility of matching based on semantically-similar phrases (paraphrases).
Thank you
Example:

Input sentence:
مذكرة من رئيس مجلس الأمن
(A memorandum by the president of the Security Council)

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

مذكرة من رئيس مجلس الأمن
ويعين مجلس الوزراء أعضاء اللجنة ويجري ...

Example

Morph. features match

Exact match