Integration of automatically-acquired multiword expressions in a hybrid machine translation system

Will Roberts and Markus Egg
Department of English and American Studies, Humboldt-Universität zu Berlin

Multimodal expression (MWE) acquisition

- Unrestricted identification of MWEs by collecting lexical co-occurrence statistics on all words in Wikipedia.
- Limited pre-processing of the text prior to MWE identification:
  - Extract plain text from the Wikipedia dumps; Segment text into sentences; tokenize and strip out URLs using regular expressions; Remove all punctuation.
  - No further processing (pos-tagging, lemmatisation, case normalisation, removal of numbers or symbols).
- Unlemmatised text may be useful for capturing the morphological and syntactic fixedness of some idiomatic MWEs (e.g., spill the beans but not spill the bean).
- Rank MWE candidates using the log-likelihood association measure.
  - Collect word frequency information using the skillt language modelling toolkit.
  - Count n-grams with n up to 3 (i.e., we treat MWEs as bigrams and trigrams).

Compositionality ranking

- Take the top 10% from each association-measure-ranked list of MWEs and re-rank these candidates in order of increasing compositionality.
- Based on Salehi et al. (2015), this makes use of word embeddings constructed using word2vec:
  - Build a vector representation for every word in the vocabulary, as well as for every MWE, using the extracted Wikipedia text.
  - Greedy string search-and-replace of all occurrences of MWEs.
  - Replace each of these with a single words-with-spaces token.
- Problem: greedy rewriting cannot handle MWEs which overlap.
- Solution: split MWEs into batches with no overlaps.
  - Each batch produces a word embedding space.
  - Compute compositionality scores, and merge batches back together.
- Compositionality score: cosine similarity of MWE vector with its constituent words (arithmetic mean).
  - Do not compute similarity with "stop words" (the 50 most frequent words in the vocabulary).

Integration into the TectoMT machine translation system (English-Spanish)

- TectoMT (Žabokrtský et al., 2008) is a hybrid machine translation system built on a pipeline model; statistical analysis phases (e.g., parsing, transfer) are interleaved with rule-based components.
- The system analyses source text up to a high (tektogrammatical) level of abstraction: a dependency graph containing only autosemantic words.

Results

- QTLeap test corpus contains 1K sentences, ca. 21K words of text from the IT domain.
- BLEU scores for translation models trained on Europarl and in-domain (1.2M sentences, 24M words) text:

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Training Types</th>
<th>Test Types</th>
<th>Training Tokens</th>
<th>Test Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl</td>
<td>In-domain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>20.24</td>
<td>26.00</td>
<td>1093</td>
<td>32,956</td>
</tr>
<tr>
<td>θ = 0.1</td>
<td>21.10</td>
<td>26.46 ***</td>
<td>5,020</td>
<td>74,015</td>
</tr>
<tr>
<td>θ = 0.2</td>
<td>21.19</td>
<td>26.43 **</td>
<td>90,133</td>
<td>2,856,015</td>
</tr>
<tr>
<td>θ = 0.3</td>
<td>26.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>θ = 0.4</td>
<td>25.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>θ = 0.5</td>
<td>19.39</td>
<td>24.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Statistical significance with respect to the baseline</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Source-only analysis of automatically acquired MWEs improves translation quality for this language pair (+0.46 BLEU points).
- The improvement is only seen for the models built with the in-domain text.
- An indication that our approach is sensitive to the domain of the training data.
- Evaluation paradigm sensitive to the compositionality of the MWEs.

Discussion

- The greatest improvements over the baseline are seen with small values of θ.
  - Including more compositional MWEs (θ > 0.1) eventually reduces BLEU scores below the baseline.
  - Composite t-nodes representing compositional MWEs likely cannot be adequately translated by single lexemes.
- Methodology introduced here is:
  - Automatic and wide-coverage, allowing construction of linguistic resources with a minimum of human effort; requires no external lexical resources or language-specific tools.
  - Language-independent.
  - Domain-independent.

References