Bilingual Embeddings for Phrase-Based Machine Translation

Will Y. Zou
Richard Socher
Daniel Cer
Christopher D. Manning
Stanford University
MT is one of the most classical and useful AI problems

Phrase-Based systems are very competitive

Classical statistical methods suffer from sparsity problems for phrase semantic equivalence

A. un cas de force majeure $\Longleftrightarrow$ case of absolute necessity (an event of) (unavoidable accident)

B. 依然故我 $\Longleftrightarrow$ persist in a stubborn manner (as before)(old)(self)

Learn Distributed Semantic Representations, with neural language models
Model Description

Combining a **Neural Language Model** with **Bilingual Constraints**

- Max-margin contrastive objective for learning word embeddings
- Obtain word alignments using the Berkeley Aligner on parallel text
- Combine both objectives to constrain word embeddings for translational equivalence
Learning of Embeddings

Curriculum Training with minibatch L-BFGS of varying band sizes:
{5k, 10k, 25k, 50k, 100k}

Train in parallel

Band size

Vocabulary

Train a large number of iterations

Entire Vocabulary

Band 1
Band 2
Band 3
Band 4

... Band N1

... Band N3

Train in parallel

Vocab. by freq. low to high
A first set of Mandarin Chinese word embeddings with 100k vocabulary (downloadable from http://ai.stanford.edu/~wzou/mt/)
Application to Stanford Phrasal System

- Phrase-table scoring in an end-to-end MT system
- Competitive BLUE baseline on NIST08 (30.01), with addition data for phrase-table extraction
- Simply average word embeddings to obtain phrase representations
- Cosine similarity is used to form an MT feature
- MERT for decoder optimization
Main Results

Word semantic similarity on SemEval 2012
NIST08 Chinese-English machine translation

<table>
<thead>
<tr>
<th>Word semantic similarity</th>
<th>BLEU score on NIST08 Chinese-English translation task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
<td><strong>Sp. Corr.</strong></td>
</tr>
<tr>
<td>Prior work (Jin and Wu, 2012)</td>
<td>5.0</td>
</tr>
<tr>
<td>Tf-idf</td>
<td></td>
</tr>
<tr>
<td>Naive tf-idf</td>
<td>41.5</td>
</tr>
<tr>
<td>Pruned tf-idf</td>
<td>46.7</td>
</tr>
<tr>
<td>Word Embeddings</td>
<td></td>
</tr>
<tr>
<td>Align-Init</td>
<td>52.9</td>
</tr>
<tr>
<td>Mono-trained</td>
<td>59.3</td>
</tr>
<tr>
<td>Biling-trained</td>
<td><strong>60.8</strong></td>
</tr>
<tr>
<td><strong>Method</strong></td>
<td><strong>BLEU</strong></td>
</tr>
<tr>
<td>Our baseline</td>
<td>30.01</td>
</tr>
<tr>
<td>Embeddings</td>
<td></td>
</tr>
<tr>
<td>Random-Init Mono-trained</td>
<td>30.09</td>
</tr>
<tr>
<td>Align-Init</td>
<td>30.31</td>
</tr>
<tr>
<td>Mono-trained</td>
<td>30.40</td>
</tr>
<tr>
<td>Biling-trained</td>
<td><strong>30.49</strong></td>
</tr>
</tbody>
</table>