Efficient Extraction of Pseudo-Parallel Sentences from Raw Monolingual Data Using Word Embeddings

Benjamin Marie and Atsushi Fujita
National Institute of Information and Communications Technology, Japan
{bmarie, atsushi.fujita}@nict.go.jp

Highlights

2-step method for sentence pair extraction:
1. filtering with sentence embeddings
2. refining with a classifier

Resource-less: extraction from raw monolingual data
⇒ no need of document-level information
⇒ no strong reliance on a lexical translation model

Fast: 12x faster than the state-of-the-art method (Tillmann+, 09)

Useful: as training data for MT
⇒ up to +1.7 BLEU in domain adaptation
⇒ significant reduction of out-of-vocabulary (OOV) words

Motivation

Bilingual sentence pair
• indispensable for MT, but costly to produce
• can be extracted automatically from monolingual corpora

Weaknesses of the previous work
• relies on document pairs to reduce the search space and to use cross-lingual IR methods (Abdul Rauf+, 11; Stafiniucu+, 12)
⇒ but document pairs are rare and difficult to collect
• relies heavily on lexical translation models (Tillmann+, 09)
⇒ assume the availability of large parallel corpora
⇒ prone to collecting sentence pairs with less OOV words

Step 1: Fast filtering with embeddings

Objective: reduce the size of the search space

Requirement: score efficiently trillions of sentence pairs

Sentence pair scoring with word embeddings:
1. train source and target word embeddings on monolingual data
2. project them in the same space (Mikolov+, 13)
3. compute sentence embeddings by averaging the word embeddings
4. score all sentence pairs by computing their cosine similarity

Return: k-best target sentences for each remaining source sentence

Step 2: refining with a classifier

Objective: rerank the k-best sentence pairs with a classifier

Requirement: characterize each pair with informative features

Features:
• sntemb: cosine similarity between sentence embeddings (step 1)
• maxalem: max. alignment between word embeddings (Kajiwara+, 16)

Training data:
• positive: small set of held-out parallel sentences (e.g., 5k sent.)
• negative: randomly paired source and target sentences

Classifier: maximum entropy

Reranking:
1. score each sentence pair with the classifier
2. rerank the sentence pairs given their score

Return: 1-best sentence pair for each remaining source sentence whose score exceeds a given threshold

SMT experiments in domain adaptation

Data
Medical translation task: EMEA (Fr→En) (Carpuat+, 12)
General-domain parallel data: Europarl (1.99M sent.)
Medical-domain monolingual data: WMT+14 (Fr-En: 1M-5M sent.)

Phrase table adaptation using extracted sentence pairs (PBSMT)

<table>
<thead>
<tr>
<th>System</th>
<th>cov. constraint</th>
<th>Fr→En BLEU</th>
<th>En→Fr BLEU</th>
<th>#OOV</th>
<th>#extracted pairs</th>
<th>speed (# pairs/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not adapted</td>
<td>25.9</td>
<td>3.134</td>
<td>23.1</td>
<td>3.099</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (Tillmann+, 09)</td>
<td>✓</td>
<td>27.2</td>
<td>2.729</td>
<td>24.7</td>
<td>2.661</td>
<td>121k</td>
</tr>
<tr>
<td>Proposed method</td>
<td>✓</td>
<td>28.6</td>
<td>1.953</td>
<td>26.4</td>
<td>1.555</td>
<td>36k</td>
</tr>
</tbody>
</table>

⇒ +1.4 (Fr→En) and +1.7 (En→Fr) BLEU points of improvement
⇒ faster extraction of more useful sentence pairs
⇒ significant reduction of the number of OOV tokens
⇒ disposal of useful sentence pairs by the coverage constraint

Analysis

Most important features:
⇒ lexprob
⇒ maxalemb

Classifier accuracy (step 2):
⇒ on truly in-domain parallel sentences: 89.6%

Conclusion & future work

Conclusion
• faster than previous work and extract more useful sentence pairs
• provide a better handling of OOV
• useful in low-resource settings by leveraging monolingual data

Future work
• speed up the extraction: lower number of dimensions for word embeddings, search approximations (e.g., LSH)
• evaluate extracted sentence pairs in more downstream tasks (phrase pair extraction, NMT, ...)

Feature set Fr→En | En→Fr
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>28.6</td>
</tr>
<tr>
<td>-lexprob</td>
<td>28.5</td>
</tr>
<tr>
<td>-maxalemb</td>
<td>29.0</td>
</tr>
<tr>
<td>-lexprob</td>
<td>28.4</td>
</tr>
<tr>
<td>-maxalemb</td>
<td>28.3</td>
</tr>
<tr>
<td>-length</td>
<td>28.9</td>
</tr>
</tbody>
</table>