Unsupervised Joint Training of Bilingual Word Embeddings

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

Summary

Key points
- Joint training directly learns bilingual word embeddings (BWE) on parallel data (cf. mapping methods)
- Unsupervised machine translation generates synthetic parallel data by translating monolingual data
- Unsupervised joint training uses synthetic parallel data for pseudo-supervision

Findings
- Largely outperforms BWE obtained through unsupervised mapping in bilingual lexicon induction tasks
- Robust to noisy synthetic parallel data and takes advantage of monolingual and bilingual contexts simultaneously

Unsupervised Joint Training

Unsupervised Mapping  
(Artetxe et al., 2018; Lample et al., 2018)

Unsupervised training for initial BWE  
Iterative refinement  
Synthetic bilingual lexicon  
Pseudo-supervised mapping for BWE  
Mapped BWE

Unsupervised Joint Training (this work)

Unsupervised machine translation  
Iterative refinement  
Synthetic parallel sentences  
Pseudo-supervised joint training for BWE  
Jointly trained BWE

Evaluation in Bilingual Lexicon Induction

Are BWE unsupervisedly and jointly trained on noisy synthetic data better than unsupervised mapped BWE?

Data
- Monolingual training data: News crawl for en-de and en-fr, Common Crawl for en-id
- Test sets: “full” Muse Wikipedia bilingual lexicons

Baseline systems: unsupervised mapping
- VECMAP: fastText word embeddings mapped without supervision
  (Artetxe et al., 2018b)

Evaluated systems
- Training data: synthetic parallel data generated with unsupervised statistical machine translation (Lample+, 2018) by translating source and/or target sentences
- BIVEC: bilingual skipgram using pre-trained word alignments (Luong+, 2015)
- SENTID: skipgram on a word/sentence-ID matrix (Levy+, 2017)

Results (acc@1)

<table>
<thead>
<tr>
<th>Method</th>
<th>Data used</th>
<th>en→de</th>
<th>de→en</th>
<th>en→fr</th>
<th>fr→en</th>
<th>en→id</th>
<th>id→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>VECMAP</td>
<td>all-all</td>
<td>42.4</td>
<td>59.0</td>
<td>67.7</td>
<td>70.0</td>
<td>58.9</td>
<td>59.5</td>
</tr>
<tr>
<td>BIVEC</td>
<td>10M-0</td>
<td>45.8</td>
<td>59.2</td>
<td>73.9</td>
<td>71.3</td>
<td>70.4</td>
<td>69.7</td>
</tr>
<tr>
<td>SENTID</td>
<td>0-10M</td>
<td>43.7</td>
<td>64.4</td>
<td>72.9</td>
<td>74.3</td>
<td>67.3</td>
<td>72.3</td>
</tr>
<tr>
<td>BIVEC</td>
<td>10M-10M</td>
<td>45.4</td>
<td>64.9</td>
<td>73.9</td>
<td>73.8</td>
<td>69.5</td>
<td>72.1</td>
</tr>
<tr>
<td>SENTID</td>
<td>10M-10M</td>
<td>45.4</td>
<td>62.1</td>
<td>74.2</td>
<td>74.0</td>
<td>69.4</td>
<td>73.0</td>
</tr>
</tbody>
</table>

- BIVEC outperforms fastText when they are trained on the same data
- Joint algorithms take advantage of noisy but bilingual contexts to monolingually improve word embeddings

Evaluation in Monolingual Word Analogy

Does unsupervised joint training improve or preserve monolingually the quality of the word embeddings (as observed for supervised bilingual skipgram)?

Settings
- Task: English word analogy (Mikolov+, 2013)
- Unsupervised systems: VECMAP (bilingual mapping), BIVEC (bilingual joint training), and fastText (monolingual)

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Data used</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VECMAP</td>
<td>239M(en)-38M(fr)</td>
<td>77.8</td>
</tr>
<tr>
<td>BIVEC</td>
<td>10M(en)-10M(synthetic)</td>
<td>65.7</td>
</tr>
<tr>
<td>fastText</td>
<td>239M(en)</td>
<td>79.1</td>
</tr>
<tr>
<td></td>
<td>10M(en)</td>
<td>64.6</td>
</tr>
</tbody>
</table>

- BIVEC outperforms fastText when they are trained on the same data

Unsupervised joint training outperforms unsupervised mapping by a large margin
- More than 10 points of improvement for en-id
- Higher accuracy when synthetic parallel data do not contain synthetic English (e.g., “10M-0” for en→de, en→fr, and en→id)
- BIVEC and SENTID perform similarly
  - pre-trained word alignments are unnecessary

⇒ Joint algorithms are robust to noise and learn better BWE for bilingual lexicon induction