BilBOWA: Fast Bilingual Distributed Representations without Word Alignments

Stephan Gouws

joint work with Yoshua Bengio and Greg Corrado

1Google Inc. and 2Université de Montréal
Introduction

• Word embedding algorithms geometrically encode *distributional lexical semantics* directly from raw text in a way that is useful for:
  
  • POS, NER, SRL, Sentiment, analysis etc.

• *Cross-lingual* embeddings generalize to > 1 language

• **Ideal:** Train model on *en* embeddings, apply to *fr, de, es, etc.*
Cross-lingual Word Embeddings
Approaches

• **Offline**: align pretrained embeddings in an offline step

• **Online** (jointly train both languages):
  • Parallel-only: only utilize parallel data
  • Mixed: utilize monolingual and parallel text
Cross-lingual Word Embeddings
Translation Matrix

Learn $W$ to transform the pre-trained English embeddings into a space where the distance between a word and its translation pair is minimized:

$$\min_W \| R^{en}W - R^{fr} \|^2$$

(Mikolov et al., 2013; Faruqui et al., 2014)
Learn $\mathbf{w}$ to transform the pre-trained English embeddings into a space where the distance between a word and its translation pair is minimized:

$$\min_{\mathbf{w}} \| \mathbf{R}^{en} \mathbf{W} - \mathbf{R}^{fr} \|^2$$

(Mikolov et al., 2013; Faruqui et al., 2014)
Cross-lingual Word Embeddings II
Using only parallel data

min(\[R\])

En parallel

\[\text{distance}\]

Fr parallel

Bilingual Auto-encoders
(Chandar et al., 2014)

BiCVM
(Hermann et al., 2014)
Cross-lingual Word Embeddings III
Online methods using mono- & bilingual data

\[ L_{en}(w|h) + \Omega_A(R) + L_{fr}(w|h) \]

(Klementiev et al., 2012; Zou et al., 2013)
## Trade-offs

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Matrix (<em>Mikolov et al. 2013</em>)</td>
<td>• FAST</td>
<td>• Assumes a global, linear, <em>one-to-one mapping</em> exists between words in 2+ languages.</td>
</tr>
<tr>
<td>Multilingual CCA (<em>Faruqui et al. 2014</em>)</td>
<td>• Simple to implement</td>
<td>• Requires accurate dictionaries</td>
</tr>
</tbody>
</table>
# Trade-offs

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
</table>
| Translation Matrix \cite{mikolov2013distributional} | • FAST  
• Simple to implement | • Assumes a global, linear, one-to-one mapping exists between words in 2+ languages.  
• Requires accurate dictionaries |
| Multilingual CCA \cite{faruqui2014multilingual}            |                                      |                                                  |
| Bilingual Auto-encoders \cite{chandar2014bilingual}     | Simple to implement (?)              | • Bag-of-words models  
• Learns more semantic than syntactic features  
• Reduced training data  
• Big domain bias |
| BiCVM \cite{hermann2014multilingual}                     | Allows arbitrary differentiable sentence composition function |                                                  |
## Trade-offs

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Matrix (Mikolov et al. 2013)</td>
<td>• FAST</td>
<td>• Assumes a global, linear, one-to-one mapping exists between words in 2+ languages.</td>
</tr>
<tr>
<td>Multilingual CCA (Faruqui et al. 2014)</td>
<td>• Simple to implement</td>
<td>• Requires accurate dictionaries</td>
</tr>
<tr>
<td>Bilingual Auto-encoders (Chandar et al. 2014)</td>
<td>Simple to implement (?)</td>
<td>• Bag-of-words models</td>
</tr>
<tr>
<td>BiCVM (Hermann et al., 2014)</td>
<td>Allows arbitrary differentiable sentence composition function</td>
<td>• Learns more semantic than syntactic features</td>
</tr>
<tr>
<td>Klementiev et al., 2012</td>
<td>Can learn fine-grained, cross-lingual syntactic/semantic features (depends on window-length)</td>
<td>• Reduced training data</td>
</tr>
<tr>
<td>Zou et al., 2013</td>
<td></td>
<td>• Big domain bias</td>
</tr>
</tbody>
</table>
## Trade-offs

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
</table>
| Translation Matrix (Mikolov et al. 2013) | • FAST  
• Simple to implement                                          | • Assumes a global, linear, one-to-one mapping exists between words in 2+ languages.  
• Requires accurate dictionaries |
| Multilingual CCA (Faruqui et al. 2014) |                                                                       |                                                       |
| Bilingual Auto-encoders (Chandar et al. 2014) | Simple to implement (?)                                              | • Bag-of-words models  
• Learns more semantic than syntactic features  
• Reduced training data  
• Big domain bias |
| BiCVM (Hermann et al., 2014)   | Allows arbitrary differentiable sentence composition function         |                                                       |
| Klementiev et al., 2012     | Can learn fine-grained, cross-lingual syntactic/semantic features (depends on window-length) | • SLOW  
• Requires word-alignments (GIZA++/Fastalign) |
| Zou et al., 2013             |                                                                       |                                                       |

This work makes cross-lingual distributed feature learning more efficient for transfer learning and translation.
BilBOWA Architecture

En monolingual

En-Fr parallel

Fr monolingual

“Cross-lingual Regularizer”: How?
We want to learn similar embeddings for translation pairs. The exact cross-lingual objective to minimize is the weighted sum over all distances of word-pairs:

\[ \Omega_\theta(A) = \sum_{i \in V^e} \sum_{j \in V^f} a_{ij} \| R^e_{[i,:]} - R^f_{[j,:]} \|_2. \]

**Main contribution:** We approximate this by sampling parallel sentences.

(Klementiev et al., 2012; Zou et al., 2013)
BilBOWA Cross-lingual Objective I

Intuition

Estimate global **alignment** statistics $P(w^e, w^f)$ from local **co-occurrence** statistics:

```latex
\approx \frac{1}{S} \sum_{s^e, s^f} S
```

Requirements:
- Requires word-level alignments
- Expensive $O(|V^e| \cdot |V^f|)$

- Requires parallel text
- Cheap $O(|s^e| \cdot |s^f|)$
BilBOWA Cross-lingual Objective II
The Approximation

$$\Omega_\theta(A) = \sum_{i \in V^e} \sum_{j \in V^f} a_{ij} \| R^e_{[i,:]} - R^f_{[j,:]} \|_2$$
BilBOWA Cross-lingual Objective II

The Approximation

$$\Omega_\theta (A) = \sum_{i \in V^e} \sum_{j \in V^f} a_{ij} \| R_{[i,:]}^e - R_{[j,:]}^f \|_2$$

$$= \ldots$$

$$\approx \| \sum_{i \in s^e} R_{[i,:]}^e - \sum_{j \in s^f} R_{[j,:]}^f \|_2$$

“BilBOWA-loss”: minimize L2-distance between BOW representations of 2 sampled parallel sentences

Much cheaper because $|s^*| << |V^*|$
Naive optimization leads to over-regularization of the frequent words. **Solution**: Subsample $w$ from bilingual sentences $\propto \Pr(w)$.
En-Fr Qualitative Analysis
En-Fr Qualitative Analysis
En-Fr Qualitative Analysis
En-Fr Qualitative Analysis
En-Fr Qualitative Analysis
En-De CLDC Experiments

Exact replication (obtained from the authors) of Klementiev et al.'s cross-language document classification (CLDC) setup:

**Goal:** Classify documents in a target language using only labelled documents in a source language.

**4 Labels:**
- **CCAT** (Corporate/Industrial),
- **ECAT** (Economics),
- **GCAT** (Government/Social), and
- **MCAT** (Markets)
En-De CLDC Experiments

Comparable or better than SOA at 3x-2,400x speedup. Trained on same data (50M words)
Comparable or better than SOA at 3x-2,400x speedup. Trained on same data (50M words)
**En-Es WMT Word Translation**

- Trained BilBOWA model on En-Es Wikipedia/Europarl data.
  - Vocabulary = 200K
  - Embedding dimension = 40
  - Window sizes in \{4, 8\}

- Exact replica of (Mikolov, Le and Sutskever, 2013):
  - Evaluated on WMT11 lexicon, translated using GTranslate
  - Top 5K-6K words as test set
En-Es WMT Word Translation

![Bar chart showing comparison of Edit Distance, Word Cooc., Translation Matrix, and BilBOWA for EnSp P@1 and EnSp P@5, and SpEn P@1 and SpEn P@5.](chart.png)
En-Es WMT Word Translation

- EnSp P@1
- EnSp P@5
- SpEn P@1
- SpEn P@5

Bar chart showing performance metrics across different scenarios:

- Edit Distance
- Word Cooc.
- Translation Matrix
- BilBOWA
Useful tricks

- Asynchronous implementation significantly speeds up training with no noticeable impact on quality

- Had to clip gradients to get to work as #dimensions grows

- Parallel subsampling improves quality for frequent words (with slight speedup)
Conclusion

- A fast hybrid cross-lingual word embedding model
- Leverages freely available monolingual data
- Uses only a small sample of sentence-aligned parallel text
- Orders of magnitude faster than other joint methods
- Improved results for
  - en-de cross-lingual document topic classification
  - en-es word translation

Code: https://github.com/gouwsmeister/bilbowa