Unsupervised Bilingual Lexicon Induction

Mareike Hartmann
Overview

• The task: Bilingual Dictionary Induction

• Approaches to it
  • Supervised
  • Semi-supervised
  • Unsupervised
The Task: Unsupervised Bilingual Dictionary Induction

Seed dictionary

- cat → gatto
- car → auto
- deep → profondo
- feline → felino

(A)

X

Y

 profondo

WX

 gatto

 auto

 cat
Generative Adversarial Networks (GANs)

Real or Fake

Discriminator

Real English embeddings

“Fake” English embeddings generated from the French embeddings

Generator

French embeddings
Possible Reasons for Failure

If unsupervised learning fails, this could be due to

- Linguistic differences resulting in unalignable embedding spaces
- Problems during model optimization
Correlation between Supervised and Unsupervised Performance

Figure 1: Correlation between P@1 scores from supervised Procrustes Analysis and unsupervised GAN-based alignment.
Translating English to English (unsupervised)

<table>
<thead>
<tr>
<th></th>
<th>SGNS</th>
<th>SVD</th>
<th>CBOW</th>
<th>Glove</th>
<th>FastText</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGNS</td>
<td>0.997</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVD</td>
<td>0.000</td>
<td>0.992</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBOW</td>
<td>0.000</td>
<td>0.000</td>
<td>0.997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glove</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.997</td>
<td></td>
</tr>
<tr>
<td>FastText</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Table 1: P@1 for unsupervised English to English alignment with embeddings trained by different embedding algorithms.