Accurate Cross-lingual Projection between Count-based Word Vectors by Exploiting Translatable Context Pairs

Shonosuke Ishiwatari*, Nobuhiro Kaji*, Naoki Yoshinaga*, Masashi Toyoda*, Masaru Kitsuregawa*

{ishiwatari, kaji, ynaga, toyoda, kitsure} @tkl.iis.u-tokyo.ac.jp

The University of Tokyo, ∗IIS, the University of Tokyo, ∗NICT, Japan, ∗NII, Japan

Overview

Problem: Word vectors in different languages are not comparable because they are learned from different corpora.

Approach: Project a vector from a language into another language space [Fung+, 98][Mikolov+, 13]

Idea: Incorporate previous approaches to realize an accurate projection between word vectors!

Previous studies

Dictionary-based approach [Fung+, 98]

Directly map count-based word vectors for each dimension by using a dictionary

( friend , amigo )
( cat , gato )

Can use relationships between context words
Only one-to-one mapping is allowed

Learning-based approach [Mikolov, 13]

Learn a linear transformation (i.e., a matrix) between predict-based word vectors

( friend , amigo )
( cat , gato )

Can find many-to-many correlation between the elements automatically
Can’t utilize the context words to learn a model

Proposed approach

• Use count-based word vectors to utilize the knowledge about context words
• The existing knowledge are obtained from
  - training set and
  - surface similarity between context words
• Weight the corresponding elements in the matrix with two bonus terms

Experiments

Data:
• Wikipedia dumps for learning word vectors
• Open Multilingual Wordnet for extracting bilingual pairs to train and test the projection
  - Most frequent 11k words for train
  - The subsequent 1k words for test

Evaluation procedure:
1. Given a word vector in the source language
2. Translate the vector into the target language
3. Choose the top-n (n = 1, 5) similar vectors in the target language
4. Check if the correct translation is included in the n vectors

Table: The accuracy of the translation

<table>
<thead>
<tr>
<th>Testset</th>
<th>[Mikolov, 13]</th>
<th>[Fung+, 98]</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Es → En</td>
<td>PIE</td>
<td>PFG</td>
<td>PIE</td>
</tr>
<tr>
<td>7.5%</td>
<td>22.0%</td>
<td>45.7%</td>
<td>61.1%</td>
</tr>
<tr>
<td>7.1%</td>
<td>18.9%</td>
<td>11.9%</td>
<td>26.1%</td>
</tr>
<tr>
<td>5.4%</td>
<td>13.8%</td>
<td>9.3%</td>
<td>22.2%</td>
</tr>
<tr>
<td>2.9%</td>
<td>11.3%</td>
<td>11.6%</td>
<td>26.8%</td>
</tr>
<tr>
<td>6.5%</td>
<td>19.1%</td>
<td>13.1%</td>
<td>22.3%</td>
</tr>
</tbody>
</table>

Top 5 translation candidates of “ニワトリ” (chicken) (en → Es)

<table>
<thead>
<tr>
<th>[Mikolov, 13]</th>
<th>[Fung+, 98]</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. kind</td>
<td>1. animal</td>
<td>1. chicken</td>
</tr>
<tr>
<td>2. call</td>
<td>2. rabbit</td>
<td>2. animal</td>
</tr>
<tr>
<td>3. frequently</td>
<td>3. eat</td>
<td>3. rabbit</td>
</tr>
<tr>
<td>4. turn</td>
<td>8. chicken</td>
<td>4. eat</td>
</tr>
<tr>
<td>5. make</td>
<td>5. fish</td>
<td>5. wild</td>
</tr>
</tbody>
</table>

Impact of the size of training data (Es → En)