Factors Influencing the Surprising Instability of Word Embeddings

Laura Wendlandt, Jonathan K. Kummerfeld, Rada Mihalcea
University of Michigan

This material is based in part upon work supported by the National Science Foundation (NSF #1344257) and the Michigan Institute for Data Science (MIDAS).
The Problem

Many common embedding algorithms have large amounts of instability.
What is Stability?

**Stability** = *percent overlap between ten nearest neighbors in an embedding space*

Stability = 40%
The Model

A ridge regression model that aims to predict the stability of a word given: (1) word properties; (2) data properties; and (3) algorithm properties.

Data:
- Europarl

Algorithms:
- word2vec skip-gram model
- GloVe
- PPMI
1. Frequency is not a major factor in stability.

**Model with frequency:**
R² score of 0.301

**Model without frequency:**
R² score of 0.301

**Model with only frequency:**
R² score of 0.008
2. Curriculum learning is important.

*Curriculum learning* = order of training data given to an algorithm
3. Stability within domains is greater than across domains.
4. POS is one of the biggest factors in stability.

<table>
<thead>
<tr>
<th>POS</th>
<th>Avg. Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeral</td>
<td>47%</td>
</tr>
<tr>
<td>Verb</td>
<td>31%</td>
</tr>
<tr>
<td>Determiner</td>
<td>31%</td>
</tr>
<tr>
<td>Adjective</td>
<td>31%</td>
</tr>
<tr>
<td>Noun</td>
<td>30%</td>
</tr>
<tr>
<td>Adverb</td>
<td>29%</td>
</tr>
<tr>
<td>Pronoun</td>
<td>29%</td>
</tr>
<tr>
<td>Conjunction</td>
<td>28%</td>
</tr>
<tr>
<td>Particle</td>
<td>26%</td>
</tr>
<tr>
<td>Adposition</td>
<td>25%</td>
</tr>
</tbody>
</table>
5. Overall, GloVe is the most stable embedding algorithm.

Word stability correlates slightly with performance on word similarity tasks.

For POS tagging using an LSTM, the LSTM compensates for instability by shifting unstable word vectors.
Final Thoughts

• *Use GloVe*
• *Learn a good curriculum for word2vec*
• *Use in-domain embeddings whenever possible*

*Download our code:*
http://lit.eecs.umich.edu/downloads.html