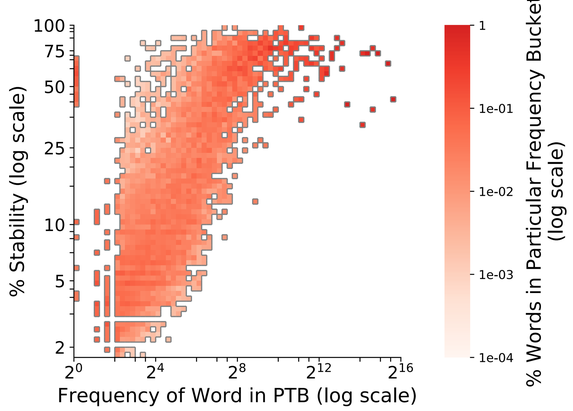
Factors Influencing the Surprising Instability of Word Embeddings

Laura Wendlandt, Jonathan K. Kummerfeld, Rada Mihalcea

University of Michigan *{wenlaura,jkummerf,mihalcea}@umich.edu*

The Problem

Many common embedding algorithms have large amounts of instability.



- Why do medium-frequency words have a huge variance in stability?
- What factors affect stability?

What is Stability?

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

$$\text{stability} = \frac{100}{|\text{words}|} \sum_{\text{words}} \frac{\text{neighbors}_0 \cap \text{neighbors}_1}{10}$$

- **neighbors$_0$** = ten words most similar to the word in embedding space 0
- **neighbors$_1$** = ten words most similar to the word in embedding space 1

Example: ***international*** *in 2 embedding spaces*

Stability = 40%

Model 1	*Model 2*
metropolitan	ballet
national	**metropolitan**
egyptian	bard
rhode	chicago
society	**national**
debut	state
folk	exhibitions
reinstallation	**society**
chairwoman	whitney
philadelphia	**rhode**

The Model

We build 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 Used
- New York Times (NYT)— six domains: US, NY, Business, Arts, Sports, All NYT
- Europarl

Algorithms Used
- *word2vec* skip-gram model
- GloVe
- PPMI

Feature	*Weight*
Lower training data position of word W	-1.52
Higher training data position of W	-1.49
Primary POS = Numeral	1.12
Primary POS = Other	-1.08
Primary POS = Punctuation mark	-1.02
Overlap between corpora vocabulary	1.01
Primary POS = Adjective	-0.92
Primary POS = Adposition	-0.92

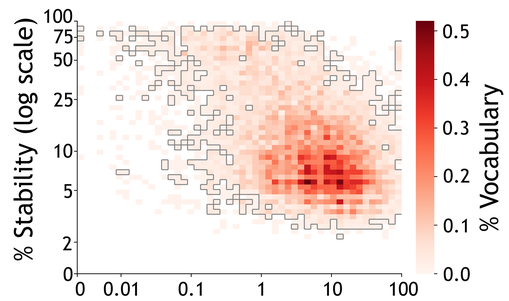
Feature	*Weight*
Do the two domains match?	0.91
Primary POS = Verb	-0.88
Primary POS = Conjunction	-0.84
Primary POS = Noun	-0.81
Primary POS = Adverb	-0.79
Do the two algorithms match?	0.78
Secondary POS = Pronoun	0.62
Primary POS = Determiner	-0.48

Lessons Learned: What Contributes to the Stability of an Embedding

1 *Curriculum learning is important.*

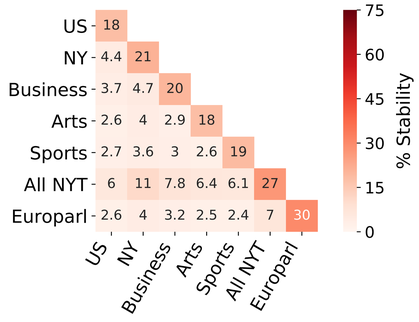
Curriculum learning = *order of training data given to an algorithm*

The top two features (by magnitude) of the regression model capture where the word first appears in the training data.



Stability of word2vec as a property of the starting word position in the training data of the PTB.

2 *Stability within domains is greater than across domains.*

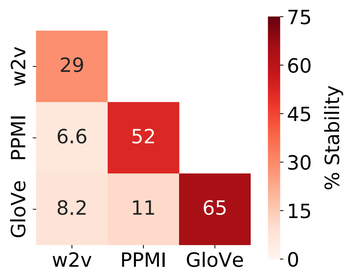


Percent stability broken down by domain.

3 *POS is one of the biggest factors in stability.*

Primary POS	*Avg. Stability*
Numeral	47%
Verb	31%
Determiner	31%
Adjective	31%
Noun	30%
Adverb	29%
Pronoun	29%
Conjunction	28%
Particle	26%
Adposition	25%

4 *Overall, GloVe is the most stable embedding algorithm.*



Percent stability broken down between algorithm (in-domain data only).

5 *Frequency is not a major factor in stability.*

Frequency does correlate with stability. However, in the presence of all of these other features, frequency becomes a minor factor.

Model with frequency:
R^2 score of 0.301

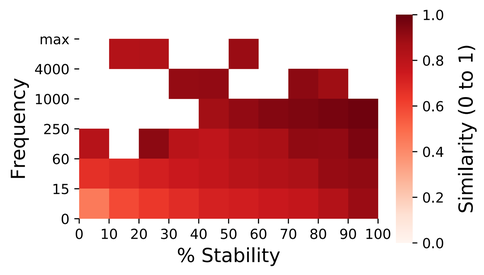
Model without frequency:
R^2 score of 0.301

Model with **only** frequency:
R^2 score of 0.008

6 *Stability affects some downstream tasks.*

Word stability correlates slightly with performance on word similarity tasks.

For POS tagging using an LSTM-based model, the LSTM compensates for instability by shifting unstable word vectors.



Word vector shift, measured as cosine similarity between initial and final vectors.

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