Understanding Word Embedding Stability Across Languages and Applications

Laura (Wendlandt) Burdick
wenlaura@umich.edu
http://laura-burdick.github.io
Words are everywhere!

- There are many ways we can use computer algorithms to do useful things with language.
Word Embeddings

- **Word embeddings**: low-dimensional vectors that capture some semantic and syntactic information about individual words
Research Questions

• Are word embeddings stable across variations in data, algorithmic parameter choices, words, and linguistic typologies?

• How does our knowledge of stability and other word embedding properties affect tasks where word embeddings are commonly used?

• How does our knowledge of stability and other word embedding properties affect our usage of embeddings?
Outline

1. Background
2. Stability in English
3. Stability in Many Languages
4. Batching & Curriculum Learning
5. Analyzing BERT
6. Conclusion
Outline

1. Background
2. Stability in English
3. Stability in Many Languages
4. Batching & Curriculum Learning
5. Analyzing BERT
6. Conclusion
Embedding Algorithms

• *Context-free output* embeddings = produce one embedding per word, regardless of word context

• *Contextualized output* embeddings = produce separate embeddings for the same word, depending on context

The word *koala* comes from…
…the *koala* is recognized worldwide…
…often miscalled the *koala* bear…

-0.7 0.55 0.63  … -0.34 0.22

The word *koala* comes from…
…the *koala* is recognized worldwide…
…often miscalled the *koala* bear…

0.5 0.67 -0.3  … -0.1 0.33

2.3 1.2 -0.5  … 0.7 -0.1
Embedding Algorithms

- Contextualized output algorithms require computational resources and data

- In some scenarios, this isn’t feasible: small datasets from digital humanities, low-resource languages

ELMo

2 wks.

$433-$1,472

262 lbs CO₂
Regression Models

Input features of interest → Fit a regression model → Metric of interest

Use the weights of the model to learn about how the features relate to our metric of interest!
Regression Models

• Ridge regression adds a regularization term

• The “goodness of fit” of a regression model is measured using $R^2$, the coefficient of determination

Image source: Fernando Wittmann on StackExchange
The Problem

- Many common embedding algorithms have large amounts of instability

2. Stability in English

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

• Percent overlap between ten nearest neighbors in two (or more) embedding spaces

philadelphia national
musatpolitan egyptian
international
tfolk rhode society
debut
reinstallation

metropolitan exhibitions ballet bard
international

national chicago
state society
whitney
rhode
What is Stability?

- Percent overlap between ten nearest neighbors in two (or more) embedding spaces

Stability = 40%
Curriculum Learning

- Curriculum learning (order of training data given to an algorithm) is important
Domains

- Stability within domains is greater than across domains

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>NY</th>
<th>Business</th>
<th>Arts</th>
<th>Sports</th>
<th>All NYT</th>
<th>Europarl</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>18</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>NY</td>
<td>21</td>
<td>5</td>
<td>20</td>
<td>18</td>
<td>19</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>Business</td>
<td>20</td>
<td>3</td>
<td>18</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>Arts</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>Sports</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>All NYT</td>
<td>6</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>Europarl</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>30</td>
</tr>
</tbody>
</table>
**Algorithms**

- Overall, GloVe is the most stable embedding algorithm
Takeaways

• Use GloVe

• Learn a good curriculum for word2vec

• Use in-domain embeddings whenever possible
Linguistic Properties

- **Key Idea**: Look at how linguistic properties of individual languages are related to stability

- **World Atlas of Linguistic Structures (WALS)**: expert-curated database of phonological, lexical, and grammatical properties for over 2,000 languages

- Does a language have a gender system?
- Does a language use suffixing?
- What is the subject, verb, object order?
Data

• Wikipedia: 40 languages

• Bible: 97 languages (at least 75% of Bible present)
Embeddings

- **Wikipedia** - 5 downsamples without replacement (100,000 sentences each), GloVe embeddings

- **Bible** - w2v with a single downsample and 5 different random seeds

![Graph showing stability in Standard Arabic with different embedding methods](image)
3. Stability in Many Languages

Embeddings

Vietnamese

Korean

% of words (log scale)

% Stability (bucketed)

Wikipedia - w2v downsamples
Wikipedia - w2v random seeds
Wikipedia - GloVe downsamples
Bible - w2v random seeds
French Bibles

- Multiple French translations (w2v with 5 random seeds)
- We except to see similar stability pattern
Regression Modeling

- Filtered to include languages and WALS properties with enough data: 37 languages, 97 properties
- Correlated WALS features grouped together
- Output: stability of all words in a language, averaged

Fit a regression model

WALS features

Average stability of a language
Model Evaluation

- $R^2$ score = 0.96 (very good)

- Leave-one-out cross-validation on all languages = average absolute error of $0.62 \pm 0.53$

- Baseline of average stability on all languages = average absolute error of $0.86 \pm 0.55$
Suffixes & Prefixes

**Strong suffixing** (inflectional morphology) or tense-aspect suffixes - 24 languages

**Weakly suffixing** (inflectional morphology) - 5 languages

**Little affixing** (inflectional morphology) - 5 languages

More affixing associated with lower stability.
Gendered Languages

3. Stability in Many Languages

No gender system associated with higher stability.
Takeaways

• We capture relationships between linguistic properties and average stability of a language

• More affixing associated with lower stability

• Languages with no gender system tend to have higher stability
Batching

- **Key Idea**: Look at different batching and curriculum learning strategies for w2v for three different tasks

![Batching Diagram](image-url)
Curriculum Learning

• **Key Idea**: Look at different batching and curriculum learning strategies for w2v for three different tasks

**Default** order of Wikipedia sentences

**Descending** order by sentence length (longest to shortest)

**Ascending** order by sentence length (shortest to longest)
Tasks

4. Batching & Curriculum Learning
Text Classification

• Smallest dataset: Real Life Deception (96 training sentences)

On the dev set, ascending curriculum with cumulative batching is best
Descending curriculum with cumulative batching is best
Takeaways

• One strategy does not perform equally well on all tasks

• Cumulative batching outperforms basic batching

• For same tasks, tuning batching and curriculum learning can substantially increase performance
Outline

1. Background
2. Stability in English
3. Stability in Many Languages
4. Batching & Curriculum Learning
5. Analyzing BERT
6. Conclusion
BERT

- Popular contextualized output embedding algorithm
Stability for BERT?

• Use paraphrases!

• Paraphrases naturally control for word semantics

• Paraphrase Database (PPDB) - word alignment, some human annotations, automatic quality score

the **goals** of the world summit

the **objectives** of the world summit
Phrase-level Embeddings

• Can BERT distinguish between two phrases that are paraphrases and two phrases that are unrelated?

• Use phrase-level embeddings
  • Average together word embeddings to get a phrase embedding
  • Take cosine similarity between two phrase embeddings
  • Compare cosine similarities to human annotations (Spearman’s correlation)
Page 41

5. Analyzing BERT

**Phrase-level Embeddings**

![Bar chart showing Spearman's rho for different word lengths and models: BERT Last Layer, w2v, and PPDB Score.]

- BERT does better with longer paraphrases.
- With longest paraphrases, BERT is comparable to PPDB score.
# Word-level Embeddings

<table>
<thead>
<tr>
<th>Aligned</th>
<th>Same</th>
<th>Different</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>adopted</strong> by the general assembly at <strong>adopted</strong> by the assembly at</td>
<td>, with a special <strong>focus</strong> on , with special <strong>emphasis</strong> on</td>
</tr>
<tr>
<td>Unaligned</td>
<td><strong>okay</strong>, so everything 's fine</td>
<td>between the canadian <strong>government</strong> and between the government of <strong>canada</strong> and</td>
</tr>
</tbody>
</table>
Word-level Embeddings

- Highest category: aligned same words
- No difference between unaligned words and aligned different words
Distance between Words

The farther away two words are in a paraphrase, the lower cosine similarity they will have.
### Polysemy

<table>
<thead>
<tr>
<th></th>
<th>1 Synset</th>
<th>5+ Synsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unaligned</td>
<td><img src="image1.png" alt="Graph for 1 Synset" /></td>
<td><img src="image2.png" alt="Graph for 5+ Synsets" /></td>
</tr>
<tr>
<td>Aligned</td>
<td><img src="image3.png" alt="Graph for Aligned" /></td>
<td><img src="image4.png" alt="Graph for Aligned" /></td>
</tr>
</tbody>
</table>

- Not a substantial difference between words with different synsets
- Aligned words more similar than unaligned words
Punctuation

Comma (178 instances)

Period (74 instances)

Question Mark (44 instances)

Dash (27 instances)

Question mark and dash used in more prescribed circumstances: question mark at end, dash at beginning.
Contextualization

• Previously, Ethayarajh [35]: BERT word embeddings are more context-specific in higher layers

• **Self-similarity**: the average cosine similarity between a word’s contextualized representations across its unique contexts
  • Self-similarity decreases, thus contextualization increases

• Instead of self-similarity, we use cosine similarity between words
Contextualization

- Decreasing similarity (increasing contextualization) for same words; same as previous work
- Increasing similarity for different words

5. Analyzing BERT
Takeaways

- BERT does a reasonable, but not perfect job controlling for semantics in paraphrases.
- BERT correctly handles polysemy in paraphrases.
- Words that are farther apart from each other in the paraphrase have lower cosine similarity scores.
- In general, paraphrased words are less contextualized than non-paraphrased words. Punctuation has highly contextual representations in BERT.
Outline

1. Background
2. Stability in English
3. Stability in Many Languages
4. Batching & Curriculum Learning
5. Analyzing BERT
6. Conclusion
Research Questions

• **Are word embeddings stable** across variations in data, algorithmic parameter choices, words, and linguistic typologies?
  • Introduced metric of stability
  • Shown that English word embedding spaces are surprisingly unstable
  • Drawn out aspects of the relationship between linguistic properties and stability for diverse world languages
  • Used paraphrases to give insight into contextualized output embedding spaces
Research Questions

• How does our knowledge of stability and other word embedding properties affect tasks where word embeddings are commonly used?
  • Showed that stability of words affects English word similarity and part-of-speech tagging (in dissertation)
  • Pinpointed linguistic properties related to instability
  • Shown how batching and curriculum learning affect performance of text classification and sentence and phrase similarity
Research Questions

• How does our knowledge of stability and other word embedding properties affect our usage of embeddings?
  • Given practical suggestions for mitigating instability in English word embeddings
  • Suggested linguistic properties as a starting point for further research on multilingual embeddings
  • Discussed tuning batching and curriculum learning for three downstream tasks
Acknowledgments

Thesis Committee:

Many, Many Others:
Drs. Bill Birmingham, Dorian Yeager, and other Grove City College professors;
Members of the LIT lab and CSE department;
Christ Church Ann Arbor; Family

Funding: NSF #1344257;
DARPA AIDA #FA8750-18-2-0019;
MIDAS
Thank you!

wenlaura@umich.edu
http://laura-burdick.github.io