Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning

Li Lucy, Jon Gauthier
Stanford NLP Group & MIT Computational Psycholinguistics Laboratory

Introduction

What are the limits of distributional meaning?
Do word embeddings produced from text alone yield sufficient knowledge about the real world?

Is there theoretical gain in modeling language learning/use as grounded or situated in more than just text?

We find systematic deficiencies in the encoding of grounded perceptual features with standard word embedding distributions.

Approach

The feature view
- Which semantic norms can be accurately predicted by distributional word embeddings?
- Learn regularized binary logic regression for each feature on word embeddings.
  - Each classifier predicts the presence/absence of a feature for each concept

The concept view
- How do deficiencies in semantic norm encoding carry over to predictions of concept similarity?
- Compare concept similarity predictions according to word embeddings and according to semantic norms

Datasets

Semantic norm datasets contain judgments of perceptual and conceptual features of natural kinds. They contain grounded knowledge about everyday objects.

Conceptual features of natural kinds.

Results

The feature view shows that, on average, word embeddings fail to encode sensory features of natural kinds. (Each point is a feature.)

The concept view shows how missing semantic features lead to mismatches in word-sense similarity predictions compared with the semantic norms and with WordNet. (Each point is a concept, color denotes the median score of the concept’s corresponding features.)

Analysis

Feature view
A bootstrap significance test shows that perceptual features are significantly worse predicted in 2 of 3 tests:

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>(functional, taxonomic)</th>
<th>(visual perceptual, other perceptual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe Common Crawl</td>
<td>(7.67%, 24.0%)</td>
<td>word2vec Google News: (7.13%, 20.6%)</td>
</tr>
<tr>
<td>GloVe Wikipedia/Gigaword</td>
<td>(-1.25%, 15.7%)</td>
<td>GloVe Wikipedia/Gigaword</td>
</tr>
</tbody>
</table>

Concept view
Feature fit deficiencies correlate with mismatches in concept similarity predictions.

See bottom graph in Results; r < 0.0196 between m(GloVe-CC, CSLB) and m(GloVe-WordNet).

Feature fit is a significant predictor of concept similarity match (correlation between distance predictions) according to post-hoc multiple regression F-tests.

Conclusion

- We find deficiencies in how word embeddings encode basic perceptual features of natural kinds.
  - These deficiencies correlate with mismatches in predictions of pairwise concept similarity.
  - These patterns appear in word embeddings sourced from different corpora and learned via different algorithms.

Can we fix these issues with more naturalistic data? Or do we need to expand our definition of meaning?

References


We use standard corpora and distributional word embedding algorithms to build vector representations of the concepts in semantic norm datasets.

Method | Training corpora
---|---
GloVe | Common Crawl

"...If we want to teach a system the true meaning of 'bumping into a wall,' we simply have to bump it into walls repeatedly..."