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

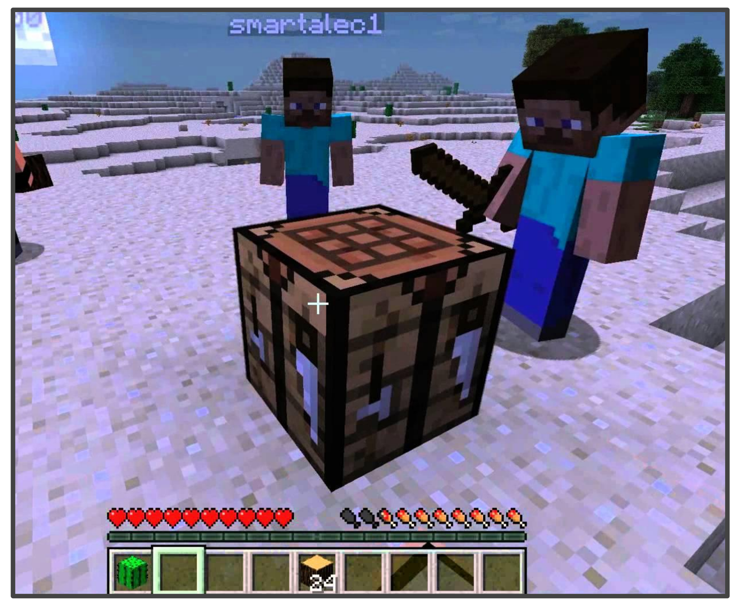
Li Lucy, Jon Gauthier
Stanford NLP Group & MIT Computational Psycholinguistics Laboratory

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.

Datasets

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

"apple"

McRae [1]	CSLB [2]
is_red	is_a_fruit
a_fruit	does_grow_on_trees
grows_on_trees	is_green
is_green	is_red
eaten_in_pies	has_pips_seeds
is_crunchy	does_grow
has_seeds	has_a_stalk_stem
is_juicy	is_circular_round
…	…

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

Method	Training corpora
GloVe [3]	Wikipedia 2014 + Gigaword 5
GloVe	Common Crawl
word2vec [4]	Google News

Approach

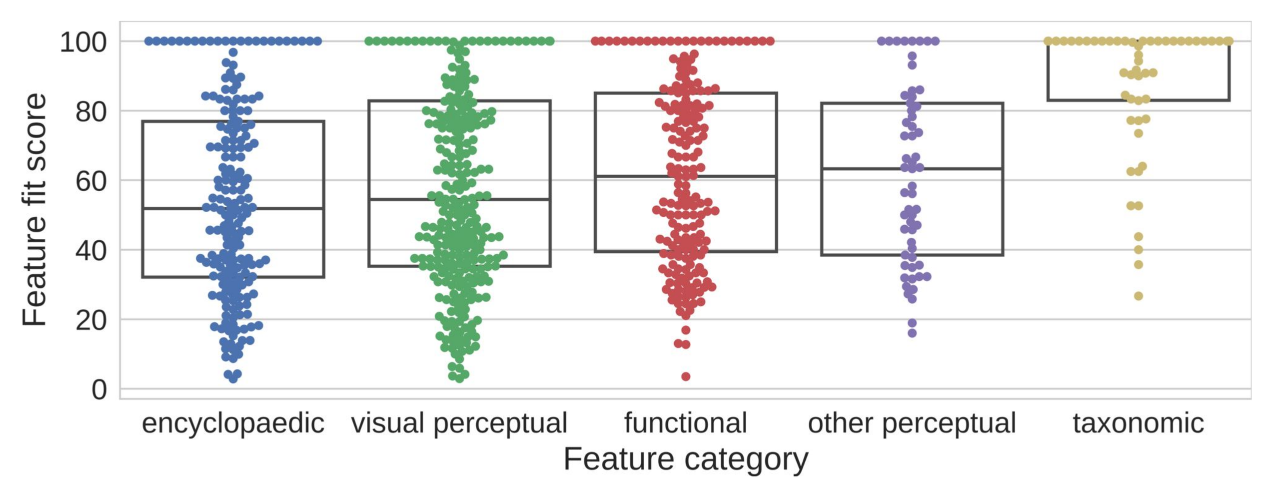
The feature view

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

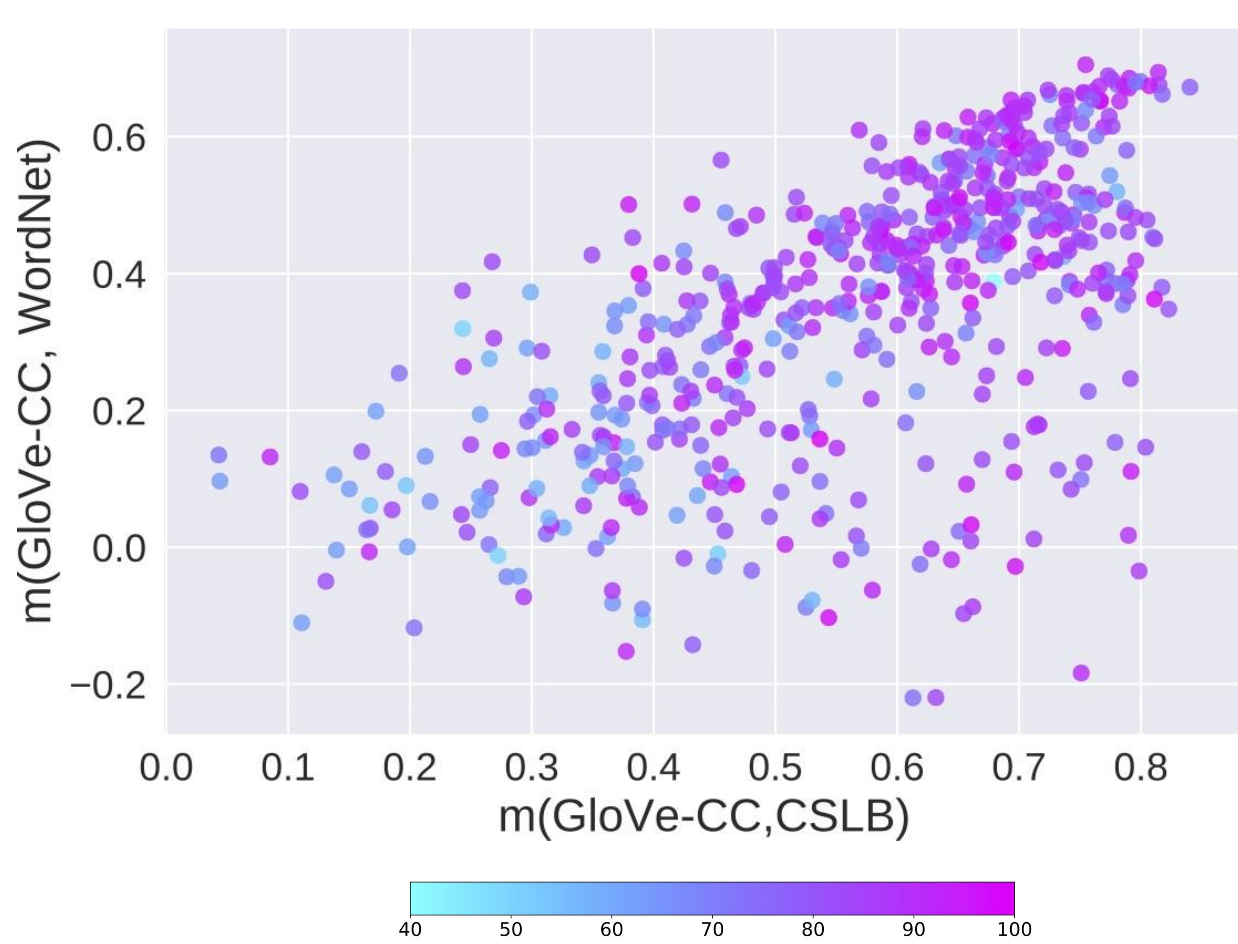
The concept view

- 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

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-word 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:

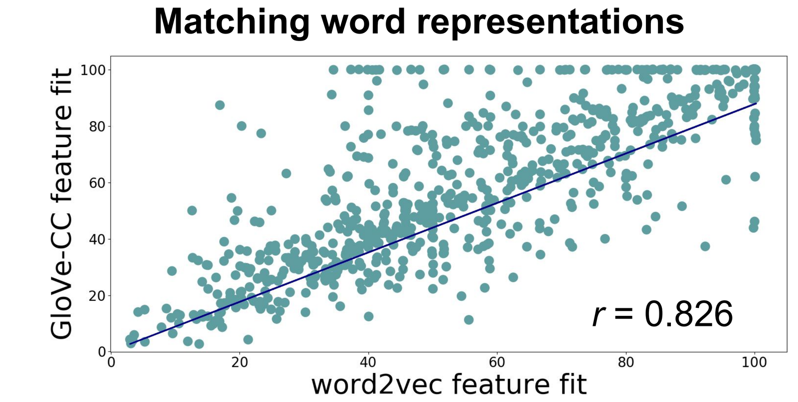
Test statistic:
(functional, taxonomic) **-** (visual perceptual, other perceptual)

95% CIs:
GloVe Common Crawl: (7.67%, 24.0%)
word2vec Google News: (7.13%, 20.6%)
GloVe Wikipedia/Gigaword: (-1.25%, 15.7%)

Table: **visual feature norms**, grouped by fit score.

< 50%	has_hands, has_a_lock, made_of_nylon, has_a_neck, is_ugly, has_a_flat_bottom, is_any_shape, is_spiky
> 50%	has_a_stone, is_slow, has_eyes, has_a_waistband, is_long, has_a_long_handle, is_colourful, has_flowers
> 90%	made_of_silk, has_whiskers, has_an_anchor, has_roots, has_pith, has_a_barrel, has_an_engine, has_sails

Matching word representations

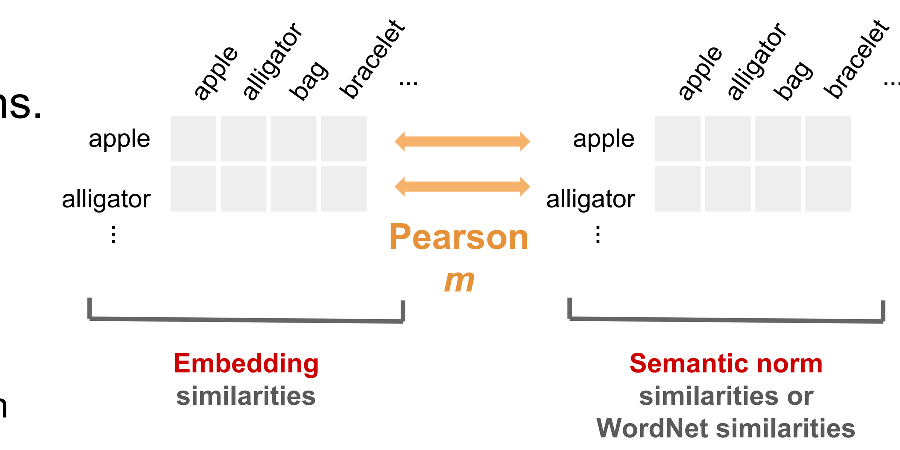


Concept view

Feature fit deficiencies correlate with mismatches in concept similarity predictions.

See bottom graph in **Results**; $r = 0.6160$ 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.

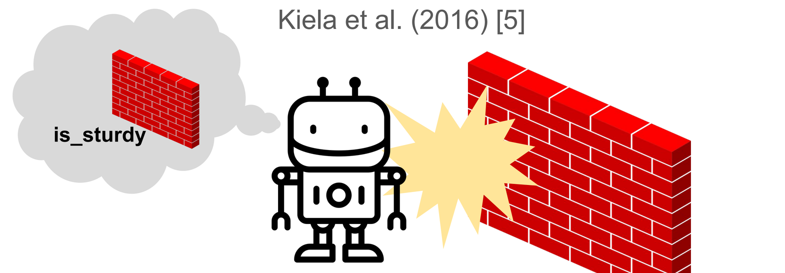


Conclusion

- 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**.

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

Kiela et al. (2016) [5]



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

References

1. Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods* 37(4):547–559.
2. Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods* 46(4):1119–1127.
3. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.
4. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532– 1543. http://www.aclweb.org/anthology/D14-1162.
5. Douwe Kiela, Luana Bulat, Anita L Vero, and Stephen Clark. 2016. Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *arXiv preprint arXiv:1610.07432* .