Semantic Primitives in Word Embeddings

Ján Pastorek

Comenius University Bratislava

Introduction

Semantic primitives are the core concepts that possibly all humans share. They cannot be defined by any other concepts, for the chain of definitions ends in them. Finding such a set would provide us with a common communication "mother language". We could use such a set to communicate ethical norms to less developed communities [1]. The list of such primes is already stable, numbering 65 in total including words such as TRUE, GOOD, NOT, YOU, etc.

Modern NLP models can capture the semantic similarity of words based on statistical co-occurrences of words. Such models create global embeddings, vectors for each word that occurs in the training where words that co-occur in similar contexts should occupy a similar place in the vector space [2]. The vector spaces produced by these models are based on co-occurrence statistics, and the models do not explicitly encode the fundamental semantic properties associated with semantic primitives.

Do the vectors corresponding to semantic primitives emerge near mathematically special regions in the vector spaces of NLP models, despite their lack of explicit encoding in those places? In other words, are the primes close to SVD singular vectors, PCA components, or K-Means cluster centers?

Methods

We have compared vectors corresponding to semantic primitives with three sets of mathematically significant, in a sense "atomic" vectors – PCA, SVD, K-Means all in a variety of settings, e.g., on reduced vocabulary to most commonly used words in English. We used three different comparison measures (Word mover's "minimum" distance similarity (WMDS), Cosine similarity (CS), Soft cosine similarity (SCS) in 6 pre-trained 300-dimensional global word embeddings models such as FastText, Conceptnet and GloVe. Lastly, we compared the results to random words for baseline.

Results

While FastText and ConceptNet performed better with WMDS, GloVe, and FastText excelled with Cosine Similarity, and GloVe was the standout performer with Squared Cosine Similarity. However, using WMDS, no effect was seen once compared to random words, thus, contradicting our hypotheses.

Conclusion

We think that the WMDS is the most reliable measure for this task since it takes into account each vector separately when comparing two sets of vectors, while other methods compare averages of sets of vectors. Moreover, we think that in some cases FastText, ConceptNet, and GloVe models captured semantic primitives near mathematically special places in the vector spaces, in this corresponding order. However, semantic primitives are not uniquely captured in the models when compared to the set of random words.

References

[1] A. Wierzbicka, "'Semantic Primitives', fifty years later," Russian Journal of Linguistics, vol. 25, no. 2, pp. 317–342, 2021.

[2] D. Jurafsky and J. H. Martin, Speech and Language Processing (3rd (draft) ed.), pp. 102 - 113, 2019.