Semantic Priming: Evaluating GloVe's Similarity Predictions Within Relational Categories

Greg Beardsley and Jacob Johnson

Spring 2022

1 Introduction

In this paper, we attempt to apply semantic Ppriming methodologies to the evaluation of GloVe vector embeddings, a tactic first suggested by Ettinger & Linzen (2016). Unlike Ettinger & Linzen, or the follow-up study performed by Auguste et al. (2017), our study does not seek to compare different embedding generation algorithms, but rather, to investigate to what degree the GloVe model captures word similarity along several semantic relations enumerated by Hutchison et al (2013).

1.1 Vector Semantics and GloVe

Vector semantics is the technology of generating numerical representations (i.e. "embeddings") for lexical items, generally to be applied to some natural language processing task. These embeddings are essentially a long list of coordinates, and can be treated as vectors or points in high-dimensional space. As such, geometric formulas (most commonly, the cosine similarity metric) can be applied to their coordinates to determine how "close" a pair of vector embeddings are, taken to indicate how related those vectors' words are. These coordinates are derived from co-occurrence statistics of words in some source corpus. The intuition behind this is that words with similar meanings will appear in similar contexts, and words with related meanings will show up in each others' contexts. The actual component values of the vectors can come directly from the co-occurrence matrix of the corpus vocabulary, or can be derived from the weights learned by some machine learning algorithm.

"GloVe" (short for "Global Vectors for word representation") is one such algorithm developed by Stanford. GloVe was observed by Ettinger & Linzen and by Auguste et al. to have a marginally higher correlation with semantic priming effects than other embedding algorithms, and for this reason it is of interest to our study. Our study generally assumes GloVe embeddings to be representative of embeddings as a whole, based on the above mentioned studies' findings of no statistically significant difference between models.

For the purposes of this study, a more detailed understanding of embeddings or GloVe in particular is not needed beyond the idea that they can be used to predict word similarity. Our study intends to further investigate to what degree these predictions indicate that the embeddings capture semantic relatedness in several dimensions.

1.2 Semantic Priming

Semantic priming is the effect where a word "primes", or facilitates quicker recall of, a semantically related word. Several models of lexical access have sprung up in the literature to account for this observation. This is a linguistic instance of a more general psychological effect, with many studies and experiments demonstrating evidence that it occurs (Swinney

et al, 1979; Perea Gotor, 1997; among others). The lexical decision task (LDT), used in experiments in the above studies, is designed to reveal this effect: participants are presented with some audio or visual stimulus, and tasked to classify it as a word or non-word. Response times and correctness are recorded. To specifically investigate priming effects, before/while a stimulus is presented to a participant, some "prime" can be briefly displayed, which might be related to the target word stimulus or not. It is expected that related primes trigger a semantic priming effect which can be detected in the response time measurements.

Further research (Hutchison et al, 2013) has investigated what specific semantic relations might trigger this semantic priming effect, yielding the following list of categories, which our study further investigates.

- Synonyms (e.g., frigid—cold)
- Antonyms (e.g., hot—cold)
- Category Coordinates (e.g., dog—animal)
- Forward phrasal associates (fpa) (e.g., help—wanted)
- Backward phrasal associates (bpa) (e.g., wanted—help)
- Perceptual properties (e.g., canary—yellow)
- Functional properties (e.g., broom—sweep)
- Script relations (e.g., restaurant—wine)
- Instrument relations (e.g., broom—floor)
- Actions (e.g., scrub—floor)
- Associated properties (e.g., deep—dark)
- Unclassified relation (e.g., mouse—cheese)

Hutchison et al. created the Semantic Priming Project (SPP), intended as a large online database of data on semantic priming effects, containing measurements from lexical decision and other semantic-priming related tasks. In the database, each prime-target word pair was classified for relation type, according to the above list.

1.3 Using Semantic Priming to Evaluate Embeddings

Vector embeddings are often remarked to work better than they actually should when it comes to improving performance on natural language processing tasks, so it is no surprise that researchers would want a way to compare them as a representation of meaning. Ettinger & Linzen report that prior work had rated embeddings using conscious word-relatedness tasks. However, since the human brain is the ultimate benchmark for language-related tasks, they proposed measuring the magnitude of semantic priming effects between words to investigate how well those correlate with similarity scores predicted by various embedding models, and found a maximal r^2 correlation of about .3. Auguste et al. extended their study by using the Spearman ranking correlation, but found a similar correlation of .25. Both of these approaches to measuring correlation give surprisingly low results. Both of these studies were based on the data from the Semantic Priming Project database, which includes classification according to a relation category.

Tangentially, the Ettinger & Linzen study noted that, across two of the categories, there is a nearly statistically-significant difference in correlation between the categories (see Figure 1), but their study did not include any further analysis. Our study aims to continue this line of research, by completing the correlation analysis within each relational category.

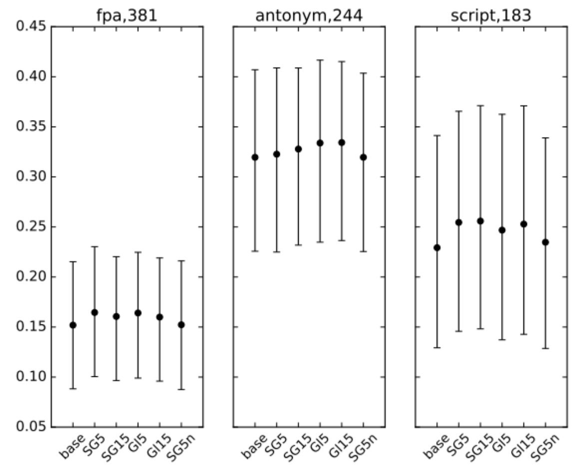


Figure 1: From Ettinger & Linzen (2016). The separate bars in each box refer to different embedding algorithms, which is not applicable to our study. The y-axis units are r^2 correlation.

2 Hypotheses

We hypothesize that embeddings do successfully capture semantic similarity, but not for the full range of semantic relations that trigger a semantic priming effect in human subjects. This hypothesis is supported by the glimpse at the analysis provided in Ettinger & Linzen (2016), and the observation that embeddings do improve computers' performance on many natural language processing tasks.

3 Experimental Design

3.1 The Semantic Priming Project

Continuing along the line of research first established by Ettinger & Linzen (2016) and Auguste et al. (2017), we will base our analysis on the data reported in the Semantic Priming Project. Originally, we considered replicating the methodology of Perea & Gotor (1997) or even the methodology of Hutchison et al. (2013), but there is no way we could begin to approach the scale of data collected by the Semantic Priming Project within a reasonable schedule. As stated by Auguste et al., such large-scale databases are becoming necessary due to the increased precision of computational models of word processing.

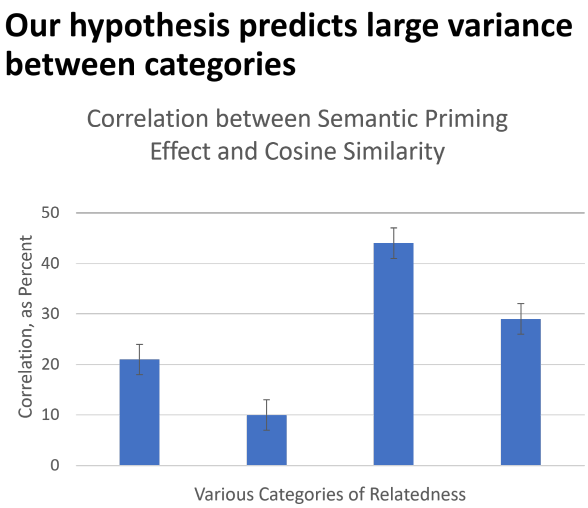
3.2 Hutchison et al. (2013) Methodology

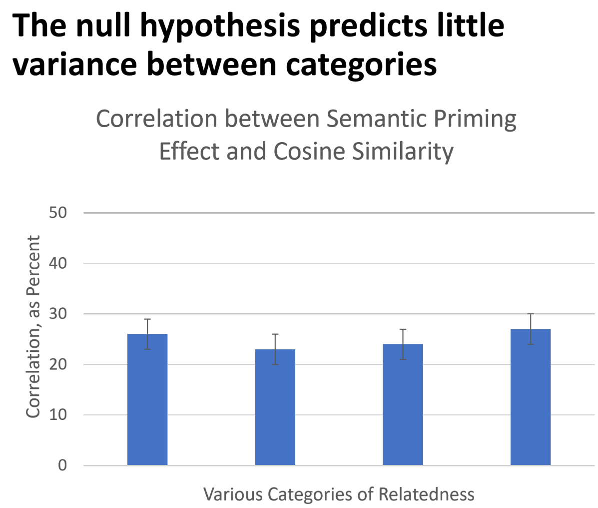
Since we will not be collecting new data for this study, we thought it would be important to summarize the methods that Hutchison et al. used in gathering the SPP data we'll be using. The Hutchison et al. project included a lexical decision task and a pronunciation task. For our purposes we will only be using data gathered from the lexical decision task. Therefore any methodology unique to the pronunciation task will not be mentioned here.

The Hutchison et al. study included 512 participants for the lexical decision task, all English speakers recruited from at or near "private and public institutions across the Midwest, Northeast, and Northwest regions of the United States". Each participant responded to 1,661 prime-target pairs during the semantic priming task. This was done in four blocks of trials across two sessions–415 or 416 trials per block and two blocks per session–with sessions no more than one week apart. Additionally, sessions included various demographic, health, and attention assessments/surveys.

The 1,661 prime-target pairs were compiled from 1,661 prime-target word pairs combined with 1,661 nonwords that were generated from the target words. Each participant saw half of the original prime-target word pairs, and half modified with nonword targets. The LDT included a fixation cross displayed for 500ms, followed by an uppercase prime word for 150ms, a blank screen for either 50 or 1050ms (depending on the LDT condition), then the lowercase target word for 3000ms. Since Hutchinson et. al. was a data collection project, it included no analyses of the data. As was intended by the project, we (as many others have done) will be analyzing the data in our study.

4 Results





As stated above, our hypothesis that embeddings do successfully capture semantic similarity within some relations predicts higher correlation in some categories, but lower correlation in other categories that are less directly tied to co-occurrence. The null hypothesis is that there is no significant difference between the results for different categories.

5 Discussion

5.1 Impact

The results of our experiment will indicate what classes of semantic similarity can meaningfully be captured by vector embedding algorithms, which also demonstrates how thorough of a representation for meaning they can be. This could aid future research into digital representations of meaning, and help predict what sorts of natural language processing tasks could most benefit from embedding representations.

5.2 Limitations

Auguste et al. (2017) point out that, in their analysis, they do not account for frequency effects or polysemy, both of which are also known to interact with lexical ecision task response times. Our similar analysis would be subject to the same limitations as theirs. Ideally, a follow up study could perform more lexical decision tasks using words that have been controlled for these variables, but it would be difficult to approach the same scale of data collection as in the Semantic Priming Project. Additionally, though we compare similarity correlations for different categories, we do not know to what degree the distributions of predicted similarities are similar across categories, nor to what degree the distributions of semantic priming effects

are similar across categories. There may be statistical methods that need to be applied to meaningfully compare observed correlations between categories.

5.3 Future Work

Though we assume GloVe embeddings to be better than, or at least representative of, embeddings as a whole, based on the observations of Ettinger & Linzen and Auguste et al., it remains possible that differences do exist between embeddings produced by different algorithms. Notably, neural network based deep contextualized embeddings debuted with ELMo shortly after Auguste et al.'s analysis, which uses sentence context when producing a word's embedding, and may have different results. Future work should compare the embeddings produced with such algorithms to those produced with more traditional GloVe algorithms. Ettinger & Linzen also suggest that future work could use neurological responses such as N400 as another implicit cognitive measure of relatedness.

6 References

Auguste, J., Rey, A., & Favre, B. (2017, September). Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In Proceedings of the 2nd workshop on evaluating vector space representations for NLP (pp. 21-26).

Ettinger, A., & Linzen, T. (2016, August). Evaluating vector space models using human semantic priming results. In Proceedings of the 1st workshop on evaluating vector-space representations for NLP (pp. 72-77).

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. Behavior research methods, 45(4), 1099-1114.

Perea, M., & Gotor, A. (1997). Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming. Cognition, 62(2), 223-240.

Swinney, D. A. (1979). Lexical access during sentence comprehension:(Re) consideration of context effects. Journal of verbal learning and verbal behavior, 18(6), 645-659.