Evaluating vector space models using human semantic priming results

Allyson Ettinger
Department of Linguistics
University of Maryland
aetting@umd.edu

Tal Linzen
LSCP & IJN, École Normale Supérieure
PSL Research University
tal.linzen@ens.fr

Abstract

Vector space models of word representation are often evaluated using human similarity ratings. Those ratings are elicited in explicit tasks and have well-known subjective biases. As an alternative, we propose evaluating vector spaces using implicit cognitive measures. We focus in particular on semantic priming, exploring the strengths and limitations of existing datasets, and propose ways in which those datasets can be improved.

1 Introduction

Vector space models of meaning (VSMs) represent the words of a vocabulary as points in a multidimensional space. These models are often evaluated by assessing the extent to which relations between pairs of word vectors mirror relations between the words that correspond to those vectors. This evaluation method requires us to select a word relation metric that can serve as ground truth, and it requires us to identify the particular types of relations that we would like our models to represent accurately.

Typical approaches to VSM evaluation use human annotations as ground truth: in particular, similarity ratings for pairs of words. Some evaluation datasets focus on similarity *per se*: *hot-scalding* would rate highly, while antonyms like *hot-cold* and associates like *hot-stove* would not (Hill et al., 2015). Others do not distinguish similarity from other types of relations: synonyms, antonyms and associates can all receive high ratings (Bruni et al., 2014).

While the distinction between similarity and relatedness is important, it represents only a preliminary step toward a more precise understanding of what we mean—and what we should mean—when we talk about relations between words. The notions of "similarity" and "relatedness" are fairly vaguely defined, and as a result human raters asked to quantify these relations must carry out some interpretations of their own with respect to the task, in order to settle upon a judgment schema and apply that schema to rate word pairs. The fact that the definition of the relation structure is left to the annotator's judgment introduces inter-annotator variability as well as potentially undesirable properties of human similarity judgments: for example, the fact that they are not symmetric (Tversky, 1977).

The subjectivity of this task, and the involvement of the conscious reasoning process needed to arrive at a rating (Batchkarov et al., 2016), raise the question: to what extent does the relation structure that emerges from such rating tasks reliably reflect the relation structure that underlies human language understanding? After all, humans process language effortlessly, and natural language comprehension does not require reasoning about how similar or related words are.

This does not mean that the brain does not perform computations reflecting relations between words—evidence suggests that such computations occur constantly in language processing, but that these computations occur on a subconscious level (Kutas and Federmeier, 2011). Fortunately, there are psycholinguistic paradigms that allow us to tap into this level of processing. If we can make use of these subconscious cognitive measures of relatedness, we may be able to continue taking advantage of humans as the source of ground truth on word relations—while avoiding the subjectivity and bias introduced by conscious rating tasks.

We propose to evaluate VSMs using semantic priming, a cognitive phenomenon understood to reflect word-level relation structure in the human brain. We show some preliminary results

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 72–77,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

exploring the ability of various VSMs to predict this measure, and discuss the potential for finer-grained differentiation between specific types of word relations. Finally, we argue that existing datasets (both explicit similarity judgments and semantic priming) are too small to meaningfully compare VSMs, and propose creating a larger semantic priming resource tailored to the needs of VSM evaluation.

2 Semantic priming

Semantic priming refers to the phenomenon in which, when performing a language task such as deciding whether a string is a word or a nonword (lexical decision), or pronouncing a word aloud (naming), humans show speeded performance if the word to which they are responding is preceded by a semantically related word (Meyer and Schvaneveldt, 1971; McNamara, 2005). For instance, response times are quicker to a word like *dog* (referred to as the "target" word) when it is preceded by a word like *cat* (referred to as the "prime"), than when it is preceded by a prime like *table*. This facilitation of the response to *dog* is taken to be an indication of the relation between *dog* and *cat*, and the magnitude of the speed-up can be interpreted as reflecting the strength of the relation.

Since priming results provide us with a human-generated quantification of relations between word pairs, without requiring participants to make conscious decisions about relatedness—the task that participants are performing is unrelated to the question of relatedness—this measure is a strong candidate for tapping into subconscious properties of word relations in the human brain.

Several studies have already shown correspondence between priming magnitude and VSM measures of relation such as cosine similarity or neighbor rank (Mandera et al., 2016; Lapesa and Evert, 2013; Jones et al., 2006; Padó and Lapata, 2007; Herdağdelen et al., 2009; McDonald and Brew, 2004). These positive results suggest that some of the implicit relation structure in the human brain is already reflected in current vector space models, and that it is in fact feasible to evaluate relation structure of VSMs by testing their ability to predict this implicit human measure.

However, to our knowledge, there has not yet been an effort to identify or tailor a priming dataset such that it is ideally suited to evaluation of VSMs. Semantic priming experiments make use of many different methodologies, and test many different types of relations between words. In selecting or constructing a priming dataset, we want to be informed about the methodologies that are best-suited to generating data for purposes of VSM evaluation, and we want in addition to have control over—or at least annotation of—the types of relations between the word pairs being tested.

3 Experimental setup

3.1 Cognitive measurements

Most previous work has modeled small priming datasets. By contrast, we follow Mandera et al. (2016) in taking advantage of the online database of the Semantic Priming Project (SPP), which compiles priming data from 768 subjects for over 6000 word pairs (Hutchison et al., 2013). This dataset's size alone is advantageous, as it potentially allows us to draw more confident conclusions about differences between models (as discussed below), and it ensures broader coverage in the vocabulary.

The SPP has two additional advantages that are relevant for our purposes. First, it contains data for four methodological variations on the semantic priming paradigm: all combinations of two tasks, lexical decision and naming, and two stimulus onset asynchronies (SOA), 200 ms and 1200 ms, which represent the amount of time between the start of the prime word and the start of the target word. We assess the usefulness of each of the methods for evaluating VSMs, in order to identify the methodological choices that generate optimal data for evaluation. A second advantage of the SPP is that it contains annotations of the relation types of the word pairs; this property can allow for finer-grained analyses that focus on relations of particular interest, as we will discuss in greater detail below.

3.2 Vector-space models

We trained four word-level VSMs for testing: skip-gram (Mikolov et al., 2013) with window sizes of 5 and 15 words (referred to as SG5 and SG15 below) and GloVe (Pennington et al., 2014) with window sizes of 5 and 15 words (Gl5 and Gl15). All models were trained on a concatenation of English Wikipedia and English GigaWord using their default parameters and dimensionality of 100. A fifth model (referred to as SG5n) was generated by adding uniform random noise $\mathcal{U}$(-2,2) to the vectors of the SG5 model, as an example of a model that we would expect to perform poorly.

3.3 Evaluation

We evaluated the VSMs by fitting linear regression models to the human response times, with cosine similarity between prime and target as the predictor of interest.[1] As a simple baseline model, we entered only word frequency as a predictor. Word frequency is widely recognized as a strong predictor of reaction time in language tasks (Rubenstein et al., 1970). While it is only one among the factors known to affect the speed of word recognition (Balota et al., 2004), it is by far the most important, and unlike factors such as word length, it is represented in many vector space models (Schnabel et al., 2015), making it all the more important to control for here.

4 Results

4.1 Cognitive measures

We first compare the four methodological variations on the semantic priming paradigm. Figure 1 shows the r^2 values, which quantify the proportion of the variance explained by the regression model. Recall that the baseline regression model ("base") contains only frequency as a predictor of response time, while the other regression models contain as predictors both frequency and cosine similarity between prime and target, as determined by each of the respective VSMs.

The greatest amount of variance is accounted for in the lexical decision task, with somewhat more variance accounted for with the 200 ms SOA. There is a more substantial margin of improvement over the frequency baseline in the 200 ms SOA, suggesting that the results of the LDT-200 ms paradigm constitute the most promising metric for assessing the extent to which VSMs reflect cognitive relation structure.

The four normally-trained VSMs (SG5, SG15, Gl5, Gl15) perform quite similarly to one another on this metric. Within those conditions in which we do see improvement over the frequency baseline—that is, primarily the lexical decision task conditions—the introduction of noise (SG5n) nullifies that improvement. This suggests that the additional variance accounted for by the four normal VSMs is indeed a reflection of their quality.

Relation	**Example pair**
Synonym	*presume, assume*
Antonym	*asleep, awake*
Forward phrasal associate	*human, being*
Script	*ambulance, emergency*
Category	*celery, carrot*
Supraordinate	*disaster, earthquake*
Instrument	*rake, leaves*
Functional property	*airplane, fly*
Backward phrasal associate	*lobe, ear*
Perceptual property	*fire, hot*
Action	*quench, thirst*

Table 1: Annotated relations in SPP

4.2 Relation types

Each word pair in the Semantic Priming Project is additionally annotated for the category of the relation between the words in the pair (see Table 1 for examples). Having access to information about the particular relations embodied by a given word pair can be quite important for maximizing the utility of our evaluation metrics, as we are likely to care about different relations depending upon the downstream task to which we intend to apply our vector representations. For instance, we may care more about faithfulness to script relations when performing document-level tasks, but care more about performance on synonym and antonym relations for word- and sentence-level tasks such as sentiment analysis and entailment.

With this in mind, we run preliminary experiments testing our VSMs as predictors of response time within the specific relation categories. In Figure 2, we show a sample of results on the per-relation level. These suggest that the spaces may vary in interesting ways, both within and between relation types. However, the small sample sizes lead to large confidence intervals; in particular, the drop in performance resulting from the addition of noise is dwarfed by the size of the error bars. As such, we cannot at this point draw firm conclusions from the results. To make conclusive use of the advantages potentially afforded by the relation annotation in the SPP, it would be necessary to collect additional relation-annotated priming data.

[1] Lapesa and Evert's (2013) result suggests that rank of the target among the vector space neighbors of the prime may model priming results more closely; we intend to experiment with this measure in future work.

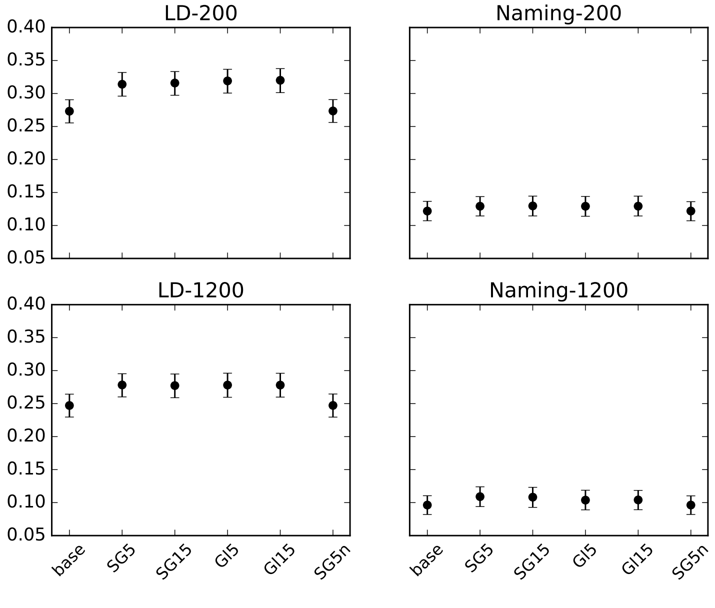

Figure 1: r^2 values for linear models fit to priming results in full SPP dataset, under different priming conditions. Baseline model ("base") contains only frequency as a predictor, while other models contain cosine values from the indicated VSMs. Error bars represent bootstrapped 95% confidence intervals.

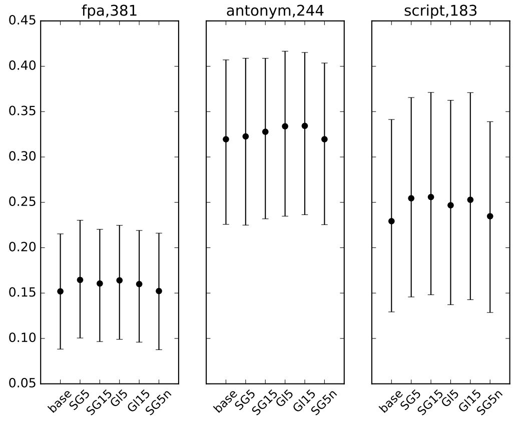

Figure 2: r^2 values for linear models fit to priming results in specific relation categories. Number of items in category is indicated in subplot title.

4.3 Similarity datasets

Finally, for the sake of comparison with conventional metrics, we include Figure 3, which shows the same baseline and vector space regression models, assessed as predictors of the ratings in the MEN (Bruni et al., 2014) and SimLex (Hill et al., 2015) datasets. Frequency appears to be a poorer predictor of explicit similarity ratings than of the implicit cognitive measures. Although there is some variation in performance between the four normally-trained VSMs, it is less straightforward to distinguish between them once we take confidence intervals into account; this issue of overlapping confidence intervals is much more pronounced with smaller datasets such as RG-65 (Rubenstein and Goodenough, 1965) and MC-30 (Miller and Charles, 1991).

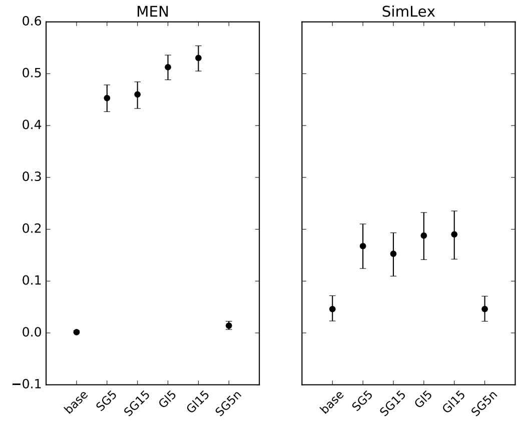

Figure 3: r^2 values, with 95% confidence intervals, for linear models fit to MEN/SimLex explicit similarity ratings.

5 Discussion

We have presented here a proposal to leverage implicit measures of relation structure in the human brain to evaluate VSMs. Such measures can sidestep the subjectivity introduced by standard similarity rating tasks, and tap more directly into the relation structure fundamental to language processing by humans.

In our exploratory results above we find, consistent with previous studies, that VSMs can predict priming beyond the variance explained by frequency alone, at least in certain cognitive measurements (in particular, lexical decision with a short SOA), suggesting that priming magnitude could be used as a VSM evaluation metric. We have also reported preliminary results taking advantage of the relation-specific annotation in the SPP. Relation-specific evaluation sets could prove valuable for finer-grained understanding of the relations captured in a given VSM. We see, however, that if we are to make statistically valid conclusions about differences between models, we must extend our dataset substantially. This could be accomplished by the same basic procedures used to build the SPP, extended to a massive scale using an online platform such as Mechanical Turk.

Finally, it may be useful to experiment with other implicit cognitive measures known to reflect relation structure. A prominent example is the N400, a neural response elicited by every word during sentence comprehension (Kutas and Federmeier, 2011). The amplitude of the N400 response is modulated by the relation of the word to its context: the worse the fit to context, the larger the N400 amplitude. As a result, the N400 is often used to study the effects of context on word processing. There is existing evidence that vector space model representations of preceding context and target words can predict N400 amplitude (Parviz et al., 2011; Ettinger et al., 2016). In future work, the N400 may therefore prove useful for assessing VSM relation structure above the word level.

Acknowledgments

This work was supported in part by an NSF Graduate Research Fellowship to Allyson Ettinger under Grant No. DGE 1322106. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF. Tal Linzen's research was supported by the European Research Council (grant ERC-2011-AdG 295810 BOOTPHON) and the Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC). This work benefited from the helpful comments of two anonymous reviewers, and from discussions with Philip Resnik, Colin Phillips, and other members of the University of Maryland language science community.

References

David A. Balota, Miachael J. Cortese, Susan D. Sergent-Marshall, Daniel H. Spieler, and Melvin J. Yap. 2004. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2):283.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the First Workshop on Evaluating Vector Space Representations for NLP*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Allyson Ettinger, Naomi H. Feldman, Philip Resnik, and Colin Phillips. 2016. Modeling N400 amplitude using vector space models of word representation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Amaç Herdağdelen, Katrin Erk, and Marco Baroni. 2009. Measuring semantic relatedness with vector space models and random walks. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 50–53.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Keith A Hutchison, David A Balota, James H Neely, Michael J Cortese, Emily R Cohen-Shikora, Chi-Shing Tse, Melvin J Yap, Jesse J Bengson, Dale Niemeyer, and Erin Buchanan. 2013. The semantic priming project. *Behavior research methods*, 45(4):1099–1114.

Michael N Jones, Walter Kintsch, and Douglas JK Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55(4):534–552.

Marta Kutas and Kara D Federmeier. 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, 62:621–647.

Gabriella Lapesa and Stefan Evert. 2013. Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013)*, pages 66–74.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2016. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*.

Scott McDonald and Chris Brew. 2004. A distributional model of semantic context effects in lexical processing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, page 17.

Timothy P McNamara. 2005. *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

D.E. Meyer and R.W. Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Mehdi Parviz, Mark Johnson, Blake Johnson, and Jon Brock. 2011. Using language models and latent semantic analysis to characterise the n400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 38–46.

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.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Herbert Rubenstein, Lonnie Garfield, and Jane A Millikan. 1970. Homographic entries in the internal lexicon. *Journal of verbal learning and verbal behavior*, 9(5):487–494.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327.