Predicting Human Similarity Judgments with Distributional Models: The Value of Word Associations

Simon De Deyne[1], **Amy Perfors**[1] **and Daniel J. Navarro**[2]
[1] Computational Cognitive Science Lab, University of Adelaide
[2] School of Psychology, University of New South Wales
simon.dedeyne@adelaide.edu.au

Abstract

To represent the meaning of a word, most models use *external* language resources, such as text corpora, to derive the distributional properties of word usage. In this study, we propose that *internal* language models, that are more closely aligned to the *mental* representations of words, can be used to derive new theoretical questions regarding the structure of the mental lexicon. A comparison with internal models also puts into perspective a number of assumptions underlying recently proposed distributional text-based models could provide important insights into cognitive science, including linguistics and artificial intelligence. We focus on word-embedding models which have been proposed to learn aspects of word meaning in a manner similar to humans and contrast them with internal language models derived from a new extensive data set of word associations. An evaluation using relatedness judgments shows that internal language models consistently outperform current state-of-the art text-based external language models. This suggests alternative approaches to represent word meaning using properties that aren't encoded in text.

1 Introduction

How is semantic information encoded? How is similarity represented in the brain? And how can we capture this information computationally? One answer to this question involves distributional lexico-semantic models, which quantify the semantic similarity between lexical items based on the distributional properties of the linguistic context in which they occur. Recent models like `word2vec` [Mikolov *et al.*, 2013] and GloVe [Pennington *et al.*, 2014], which rely on external corpora as the source of data, increasingly appear to capture word meaning in ways that ever-more-closely resemble human representations. For instance, these models show systematic improvements over previous work in key benchmarks such as human similarity judgments of word pairs [Baroni *et al.*, 2014]. The strong performance of these models has also suggested to cognitive scientists that the learning mechanisms they embody might resemble how humans learn the meaning of some words [Mandera *et al.*, 2017].

In this study we show that using word-association data instead of corpus data improves performance substantially above the current state-of-the-art. We suggest that this is because data-intensive distributional models like `word2vec`, formidable though they are, may not capture word representations the way the average adult language speaker does. Their enormous, high-quality input data enables them to mimic human behavior, but they do relatively poorly compared to performance based on data that more accurately captures people's true representations of meaning.

The distinction between using text corpora or word association data maps between External language models (E-language) and Internal language models (I-language) [Taylor, 2012]. An E-language model, like `word2vec`, treats language as an "external" object consisting of the all utterances made in a speech-community. An I-language model sees language as the body of knowledge residing in the brains of its speakers. Largely due to the easy availability of high-quality external corpora – for instance, there are over one trillion words in the Google n-gram corpus [Michel *et al.*, 2011] – computational linguists have traditionally focused on E-language models. Whether a similar distributional approach based on I-language might also be useful has received relatively less attention. One explanation could be purely on the basis of practical arguments, as it's not clear whether appropriate I-language resources are available. This paper fills that gap, by introducing an approximation of I-language using a new database of word associations considerably larger than previous ones and conducting a direct comparison of how both kinds of approaches predict human similarity judgments. It is valuable not just in demonstrating that models based on I-language greatly improve their performance. It also suggests that when people judge similarity, they may be relying more on networks of semantic associations than on statistics calculated from the distributional patterns of the words they hear. The structure of this paper is as follows. In Part 2 we describe the origin and nature of the data for the E-language source (text corpora) and I-language source (word association data). Part 3 describes the distributional models, part 4 describes the multiple human similarity and relatedness judgments that each model and data source will be used to predict. Part 5, the results, demonstrates that models based on I-language consistently perform substantially better than the same model based on E-language.

2 Data Sources

2.1 A Varied E-language Model Text Corpus

Our aim was to combine corpora that would provide us with a fairly balanced set of texts that is representative of the sort of language a person experiences during a lifetime – including both formal and informal language as well as spoken and written language.

The corpus consisted of several subcorpora including English movie subtitles, contemporary English fiction, newspaper articles, spoken text and SimpleWiki and are described in [De Deyne *et al.*, 2016b]. The resulting corpus consisted of 2.16 billion tokens and 4.17 million types. We further excluded words that did not occur at least 300 times, retaining 65,632 unique word types. This cut-off is larger than previous approaches using count models and word embedding models but allowed us to reduce the memory requirements for the count model we introduce later and to make sure that words in the evaluation sets were at least as frequent as the words in the association study for which we collected 300 responses. Altogether, this corpus was constructed to be generous in terms of the quality and quantity of items so that models incorporating it would perform similarly to the existing state-of-the-art.

2.2 A Novel Word Association Dataset for I-language Models

One of the shortcomings with previous word association studies is that they only include the strongest associations because only a single response is generated for each cue word. For example, in the case of *umbrella*, most participants would respond *rain*, which prevents the inclusion of weaker links. A better way to include weaker associates as well is by using a continued procedure where multiple responses for each cue word were collected. Extending the response set to include weaker responses and including enough cue words to capture most words used in daily languages motivated us to set up a new large-scale study. The current data are collected as part of the Small World of Words project (`https://smallworldofwords.org/`), an ongoing effort to map the mental lexicon in various languages. Each participant was given a short list of cue words (between 15 and 20 words) and asked to generate three different responses to each cue. To avoid chaining responses, the instructions stressed to only give a response to the cue word. If a word was unknown or no secondary or tertiary response could be given, the participants were able to indicate this Additional details on the procedure are available in [De Deyne *et al.*, 2013].

The results reported here are based on 10,021 cue words for which at least 300 responses have been collected (100 primary, 100 secondary and 100 tertiary) for every cue. The study was presented as an online crowd sourced project in which fluent English speakers volunteered to participate. The responses were based on over 85,496 participants. In line with previous work, we constructed a semantic graph from these data. This graph closely resembles the bag-of-words count models but represented as a graph makes it possible to consider the spreading activation discussed in the next section. A graph $\mathbf{G}$ was constructed by only including responses that also occurred as a cue word. This converted the bimodal cue $\times$ response graph to a unimodal cue $\times$ response graph. In this weighted graph $\mathbf{G}$, g_{ij} counts the number of times that word j is given as an associate of word i. We extracted the largest strongly connected component by only keeping those cues that were also given at least once as a response. This way all words can be reached by both in- and out-going links. The resulting graph consists of 10,014 nodes and the number of different word types each word is connected to (i.e., its out-degree of) is 92. As expected, the graph is also very sparse: only 0.92% of words are connected (i.e., $\mathbf{G}$ has 0.92% non-zero entries).

3 Models

We consider four different models in this paper, two E-language models estimated from the text corpora, and two I-language models that use word association data. In both cases, one model is a simple count based model and the other aims to exploit additional structure of the sparse input data.

3.1 Count Based Model for Text Corpora

Count models of text corpus data use a simple representation: they track how many times a pair of words co-occur in a document or sentence. For our analyses we used a symmetric dynamic window that linearly weighted words as a function of the distance between them. The resulting co-occurrence frequencies were transformed using the positive point-wise mutual information (PMI^{+}), given the evidence that this measure performs well in count models and combined it with a discount factor in order to prevent very rare words from biasing the results [Levy *et al.*, 2015].

3.2 Predicting Structure from Text Corpora using Word Embeddings

An alternative approach to representing text corpora is to apply a lexico-semantic model that aims to extract the latent semantic structure embedded in the text corpus by learning to predict words from context. We focused on the word embeddings derived from the neural network approach in *word2vec* [Mikolov *et al.*, 2013; Levy *et al.*, 2015], using a continuous bag of words (CBOW) architecture in which the model is given the surrounding context for a word (i.e., the other words in a sliding window) and is trained to predict that word.

Based on previous work [Baroni *et al.*, 2014; Mandera *et al.*, 2017; Levy *et al.*, 2015] the following settings were used: a negative sampling value of 10, and a down-sampling rate of very frequent terms of 1e-5. We considered window sizes between 2 to 10, and fitted models with between 100 and 500 dimensions with steps of 100. We will focus on the best fitting hyper-parameter values but note that the differences for other values were rather small.

3.3 Count Based Model for Word Associations

In an E-language model, the goal is to characterize the linguistic contents of a text corpus, whereas an I-language model aims to capture the mental representation that a human speaker might employ. The difference between these two goals motivates a difference in the kinds of data that one

might use (e.g., text corpora versus word associations) but there are commonalities between the two approaches. For example, there is evidence that the relationship between (observed) word association frequency and (latent) associative strength is nonlinear [Deese, 1965], an observation that suggests the PMI$^+$ measure might be reasonably successful as a simple count model for association strength. With that in mind our first model is a simple PMI$^+$ measure using the word association frequency as the input.

3.4 A Spreading Activation Approach to Semantic Structure

While the PMI$^+$ model captures the semantic information in the raw word association data, it does not attempt to capture any deeper semantic structure that these data encode. We use word association data to construct a network that connects associated words, and model semantic similarity using denser distributions derived from a *random walk* defined over this network [De Deyne *et al.*, 2016b]. The intuitive idea is that when a word is presented it activates the corresponding node in the graph, and starts a random walk (or many such walks) through the graph, activating nodes that the walk passes through. If there are many short paths that connect two nodes, then it is easy for a random walk through the graph to start at one node and end at the other, and the words are deemed to be more similar as a consequence.

To implement this idea we first normalize the word association matrix such that each row sums to 1, thus converting it to a transition matrix $\mathbf{P}$. In the limit, where we consider paths of arbitrarily long length through the following expression [Newman, 2010]:

$$\mathbf{G_{rw}} = \sum_{r=0}^{\infty}(\alpha\mathbf{P})^r = (\mathbf{I} - \alpha\mathbf{P})^{-1}$$

Finally, we apply the PMI$^+$ weighing function to $\mathbf{G_{rw}}$ reduces the frequency bias introduced by this type of walk [Newman, 2010] and also keeps the graph sparse.

To see how this spreading activation mechanism can be very powerful, consider the word *tiger*. Before applying spreading activation its meaning vector consists of 92 different association responses. When we apply the spreading activation measure we uncover nearly 559 new associations which ordered by their weights included *zebra*, *cheetah*, *claws*, *cougar* and *carnivore*, all of which seem meaningfully related to *tiger* but were not among the responses when *tiger* was presented as a cue word.

4 Comparing Model Predictions to Human Judgments

The data sets used to evaluate the models broadly fall into one of two classes. Two of the studies asked participants to judge the *similarity* between words, namely the WordSim-353A similarity data set [Agirre *et al.*, 2009] and the SimLex-999 data [Hill *et al.*, 2016]. In the remaining studies participants were asked to judge relatedness. These include the WordSim-353 relatedness data set [Agirre *et al.*, 2009], the MEN data [Bruni *et al.*, 2012], the Radinsky2011 data [Radinsky *et al.*, 2011], the popular Rubenstein and Goodenough (RG1965) data [Rubenstein and Goodenough, 1965] and the MTURK-771 data [Halawi *et al.*, 2012].

In addition to these data sets, we include data from a relatedness judgment task based on triadic comparisons using a procedure introduced in [De Deyne *et al.*, 2016a]. In this task, participants are asked to select the most related pair out of a set of three English nouns. An advantage of this task is that the third word acts as a context, which makes judgments less ambiguous. Critically, the triads were constructed by choosing words largely at random from the English word association data set. The only constraints were that the words in a triad had to be roughly matched on judged concreteness and word frequency. This was done to avoid simple heuristics such as grouping abstract or common words together. The consequence of this procedure is that the triads tended to consist of words that are only weakly related to each other, such as BRANCH - ROCKET - SHEET or CLOUD - TENNIS - SURGEON, and it is for this reason it is referred to as the "remote triads task". A total set of 100 triads was constructed this way and judgments were collected for 40 native English speakers. Because the words are chosen at random, we expect that a large portion of them will only share a small number of features in the count models which poses an ideal scenario to test how prediction and spreading activation approaches induce additional structure from this sparse input.

All four models represent word meanings as a semantic vector, and we used the cosine similarity measure in all cases. Only word pairs that were present in the text corpus and the word association data were included. As shown in Table 1 (columns 2 and 3), most words were retained. For the triads task model predictions were obtained by normalizing the similarities between the three words in each triad and correlating them with the frequencies of the choice preferences.

5 Results

The best performing parameters were a window size of 3 for the corpus count model, and a window size of 7 and 400 dimensions for *word2vec*, although the findings for other window sizes and dimensions were quite similar. The word association count model is based on $\mathbf{G_{123}}$ and has no free parameters, whereas for the random walk model we used a parameter value of α = 0.75, similar to previous studies [De Deyne *et al.*, 2016a]. Table 1 shows the performance of all models, and it is clear that the I-language models substantially outperform the E-language models in almost every case. It is also clear that extracting structure helps: *word2vec* generally outperformed the corpus count model, and the random walk model outperformed the word association count model. For the E-language models the magnitude of this effect was slightly smaller than reported elsewhere [Baroni *et al.*, 2014; Mandera *et al.*, 2017]. Surprisingly, the count model outperformed *word2vec* on the remote triad task which questions how human-like learning is in word2vec.

Apart from the results reported here, we also piloted E-language models that used different frequency cut-off values and used embedding vectors that have been previously published elsewhere, showing very similar results [De Deyne *et al.*, 2016b].

Table 1: Spearman rank order correlations between relatedness and model predictions for all four models described in the text.

			Text Corpus		**Word Associations**	
Data set	*n*	*n(overlap)*	Count	*word2vec*	Count	Random Walk
WordSim-353A	252	207	.67	.70	.77	.82
WordSim-353B	203	175	.74	.79	.84	.87
MTURK-771	771	678	.67	.71	.81	.83
SimLex-999	998	927	.37	.43	.70	.68
Radinsky2011	287	137	.75	.78	.74	.79
RG1965	65	52	.78	.83	.93	.95
MEN	3000	2611	.75	.79	.85	.87
Remote Triads	300	300	.65	.52	.62	.74
mean			.67	.69	.78	.82

6 Discussion

The goal of this study was to compare two kinds of semantic models: "I-language" models that encode mental representations, and "E-language" models that encode lexical contingencies. In one respect the superior performance of the I-language models is unsurprising: the training data directly reflect human mental representations, and as such *should* be more strongly linked to human semantic judgments. On the other hand, the I-language models were trained on a *much* smaller data set than the E-language models, with an average of 260 words contributing to the distributional representation of each word. Given this, it is worth considering the broader implications of the findings.

First there is the issue of the role of learning. Previous work has argued that the *word2vec* model is more cognitively plausible than count models due to its similarity to models of classical conditioning [Mandera *et al.*, 2017]. This is contrasted with more statistical approaches such as Latent Semantic Analysis [Landauer and Dumais, 1997] and topic models [Griffiths *et al.*, 2007]. However, it is not clear that this holds up when we find very little difference in performance between count models and *word2vec*, or previous work arguing that word embedding models perform an implicit matrix factorization [Levy and Goldberg, 2014]. Perhaps more importantly, there is something strange about the claim that E-language models are cognitively plausible when the data sets upon which they are trained are as large as they are. If purely text based models are intended to stand as models for how humans acquire semantic structure, then they should be trained on a corpus small enough that it plausibly represents the language exposure of the young adults who participated in the benchmark tasks. If billions of tokens are required to produce adequate predictions while still being unable to match the performance of simple I-language models, it is not clear what claims can be made about human language acquisition.

Next there is the issue of the nature of the representations. To understand why I-language models perform so well using limited amounts of data and to set a direction of how to improve E-language models, it is useful to consider what kind of semantic information word associations capture representations that cannot be fully reduced to the distributional properties of the E-language environment. Previous attempts to predict word associations from E-language have had limited success [Griffiths *et al.*, 2007]. E-language typically only predicts the strongest associate in the minority of cases and does even worse in predicting non-primary responses. Why is this? At least part of it is that E-language has the structure it does because people are using it to communicate to each other; it is not simply a reflection of their mental representations. For instance, *yellow* is a very strong associate of *banana*, but the two words co-occur relatively infrequently since most bananas are yellow. As a result, modifying the word *banana* with *yellow* is uninformative, so most people leave it out when talking. Many of the divergences between the distributions of words in external language and the strength of internal associations may occur because so much of E-language is shaped by pragmatic and communicative considerations such as these. Other evidence suggests that mental representations, as reflected in word associations, are shaped by far more than the distributional properties of the E-language. For instance, fMRI measures reveal the activation of imagery-related areas during word association tasks [Simmons *et al.*, 2008]. This suggests that a compact set of I-language features (perceptual or other) can provide us with valuable pointers towards further refining existing E-language models and NLP applications build from them. For example, recent E-language models have taken a multimodal approach by enhancing language sources with visual representations [Lazaridou *et al.*, 2015]. To illustrate, [Bruni *et al.*, 2012] evaluated word embeddings combined with features extracted from images on the MEN dataset and found a correlation $r = .78$ for the best performing model, which is considerably lower than current results for the E-language and especially the I-language models. Similarly, more recent work of our own shows that multimodal models that take into account both perceptual and emotional internal states provide substantial improvements for E-language models but only marginally improve I-language models, supporting the idea that the latter already provide a symbolic representation of meaning that encodes perceptual and emotional information beyond what's accessible from language alone.

Acknowledgments

This research was supported through ARC grants DE140101749 to SDD, DE120102378 to AP, and FT110100431 to DJN. This paper is an abridged version of [De Deyne *et al.*, 2016b], presented at COLING-2016.

References

[Agirre *et al.*, 2009] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. ACL, 2009.

[Baroni *et al.*, 2014] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–247, 2014.

[Bruni *et al.*, 2012] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics, 2012.

[De Deyne *et al.*, 2013] Simon De Deyne, Daniel J Navarro, and Gert Storms. Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations. *Behavior Research Methods*, 45:480–498, 2013.

[De Deyne *et al.*, 2016a] Simon De Deyne, Daniel J Navarro, Amy Perfors, and Gert Storms. Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, 145:1228–1254, 2016.

[De Deyne *et al.*, 2016b] Simon De Deyne, Amy Perfors, and Daniel J. Navarro. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of the 26th International Conference on Computational Linguistics*. ACL, 2016.

[Deese, 1965] James Deese. *The structure of associations in language and thought*. Johns Hopkins Press, 1965.

[Griffiths *et al.*, 2007] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.

[Halawi *et al.*, 2012] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM, 2012.

[Hill *et al.*, 2016] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41:665–695, 2016.

[Landauer and Dumais, 1997] Tom K. Landauer and Susan T. Dumais. A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[Lazaridou *et al.*, 2015] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.

[Levy and Goldberg, 2014] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014.

[Levy *et al.*, 2015] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[Mandera *et al.*, 2017] Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 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*, 92:57–78, 2017.

[Michel *et al.*, 2011] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182, 2011.

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

[Newman, 2010] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[Radinsky *et al.*, 2011] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM, 2011.

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

[Simmons *et al.*, 2008] William K. Simmons, Stephan B. Hamann, Carla N. Harenski, Xiaoping P. Hu, and Lawrence W. Barsalou. fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology - Paris*, 102:106–119, 2008.

[Taylor, 2012] John R. Taylor. *The mental corpus: How language is represented in the mind*. Oxford University Press, 2012.