Predicting human similarity judgments with distributional models: The value of word associations

Simon De Deyne and Amy Perfors
Computational Cognitive Science Lab
School of Psychology
University of Adelaide
simon.dedeyne@adelaide.edu.au
amy.perfors@adelaide.edu.au

Daniel J Navarro
School of Psychology
University of New South Wales
dan.navarro@unsw.edu.au

Abstract

Most distributional lexico-semantic models derive their representations based on *external* language resources such as text corpora. In this study, we propose that *internal* language models, that are more closely aligned to the mental representations of words could provide important insights into cognitive science, including linguistics. Doing so allows us to reflect upon theoretical questions regarding the structure of the mental lexicon, and also puts into perspective a number of assumptions underlying recently proposed distributional text-based models. In particular, we focus on word-embedding models which have been proposed to learn aspects of word meaning in a manner similar to humans. These are contrasted with internal language models derived from a new extensive data set of word associations. Using relatedness and similarity judgments we evaluate these models and find that the word-association-based internal language models consistently outperform current state-of-the art text-based external language models, often with a large margin. These results are not just a performance improvement; they also have implications for our understanding of how distributional knowledge is used by people.

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 their distributional properties in large samples of data. 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., in press).

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 onto the distinction made by Taylor (2012) between External language models (E-language) and Internal language models (I-language). 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

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1861–1870, Osaka, Japan, December 11-17 2016.

et al., 2011) – computational linguists have traditionally focused on E-language models (Bullinaria and Levy, 2007; Baroni et al., 2014; Levy et al., 2015). 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.

Why should we expect to see (and why *do* we see) such improvements when the models use word associations data rather than high-quality large-scale text corpora data? After all, word association models generally incorporate far less data. Moreover, one might presume that word associations are themselves simply derived from the distribution of words in the external language: in that case, one would expect them to be an inferior and noisy measure.

However, several strands of research support the idea that 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; Michelbacher et al., 2007; Wettler et al., 2005). 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, the word "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. There is also evidence that meaning representations in the brain, as reflected in word associations, are shaped by far more than the distributional properties of the E-language. For instance, fMRI measures reveal that imagery-related areas like the precuneus are activated during word association tasks (Simmons et al., 2008).

The structure of this paper is as follows. In Part 2 we describe the origin and nature of the data we are using as the E-language source (text corpora) and I-language source (word association data). Part 3 describes the distributional models which we will apply to each data source, while 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

The central comparison in this paper is between model performance on E-language vs I-language data. The increasing number of online resources from which text can be extracted means that obtaining a representative E-language has become more straightforward. Furthermore, better balanced corpora that easily surpass the knowledge of the average human are readily available. We derived our E-language data based on four different kinds of existing text-based corpora, as described below.

Free word association data are used as the I-language data. Although there are certainly other possibilities, word association data are advantageous as they appear to tap directly into mental representations (Deese, 1965; McRae et al., 2012; Szalay and Deese, 1978). Moreover, we shall show that a new procedure and larger data sets address important shortcomings from previous work that relied on a relatively small number of cues words for which only a single association response was asked from the participant (Kiss et al., 1973; Nelson et al., 2004).

2.1 A text corpus to train E-language models

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. Four different corpora were combined.

1. Subtitles for English movies between 1970 and 2016 extracted from the OpenSubtitle corpus as described in Tiedemann (2012). Subtitle corpora have been frequently used in cognitive science because they capture daily language better than extremely large written corpora like the Google n-gram corpus (Brysbaert et al., 2011).

2. The Corpus of Contemporary English (COCA), as described in Davies (1990 present). It consists of a balanced set of formal and informal language including fiction, newspaper articles and spoken texts. We excluded the sub-corpus for academic texts.

3. The Global Web-Based English corpus (GlowBE), as described in Davies (2013). We included the sub-corpora of British, American, Canadian and Australian texts.

4. SimpleWiki, which presents knowledge that is likely available to the average person (18 million tokens in comparison to the 2.9 billion words in the full English Wikipedia).

Altogether, in compiling these corpora we aimed 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. For similar reasons, we used word-forms rather than lemmas: this matches previous work and provides the best possible match with the stimuli in the human benchmarks. The resulting corpus consisted of 2.16 billion tokens and 4.17 million types. Each sentence was uncased and stop words were removed. 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. Moreover, we piloted different cut-offs and found it didn't affect our findings.

2.2 A novel word association dataset for I-language models

One of the shortcomings with previous word association studies of considerable size like the Edinburgh Association Thesaurus (Kiss et al., 1973) or the University of South Florida norms (Nelson et al., 2004) is that they only include the strongest associations (Aitchison, 2012) 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 (Szalay and Deese, 1978). 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, an ongoing effort to map the mental lexicon in various languages [1]. 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 of which 82% were native speakers. Responses indicated as

[1] The word association task and details of the project can be accessed at `https://smallworldofwords.org/`. A paper describing an extension of these norms including over 12,000 cues is currently in preparation and the data will be made available on the same website.

unknown (1.17% of cue words) or missing, because a participant could not think of any secondary or tertiary associations (4.15% of responses), were excluded. 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, which retains 84% of the original data consisting of all responses. The average number of tokens per word is 267 and the number of different word types each word is connected to (i.e., its out-degree of) is 92, ranging from 12 (for the word *done*) to 169 (for *control*). 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). Throughout the text we will refer to this graph as $\mathbf{G_{123}}$, since it incorporates all three responses given by participants.

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 the structure of the 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 applied a sliding window at the sentence level similar to Pennington et al. (2014). Specifically, we used a symmetric dynamic window that linearly weighted words as a function of the distance between them (Pennington et al., 2014). 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 (Bullinaria and Levy, 2007; Levy et al., 2015). In particular, we follow Levy et al. (2015) in applying a discount factor in order to prevent very rare words from biasing the results (see their Equation 3), and unless otherwise stated we used the same discount factor (0.75) that they did.

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.

For our analyses, we used the *gensim* implementation of *word2vec* (Řehůřek and Sojka, 2010). Based on previous work (Baroni et al., 2014; Mandera et al., in press) the following settings were used: a negative sampling value of 10, and a down-sampling rate of very frequent terms of 1e-5. Additionally, the following hyper-parameters were manipulated, using the values reported by previous work (Levy et al., 2015) as a starting point. 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 for the purposes of robustness we will also examine the performance of models using previously published semantic vectors.

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.[2]

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. Inspired by classic work in human semantic memory by Collins and Loftus (1975), 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, similar to the Katz index (Katz, 1953). 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 $\mathbf{P}$. Then, in order to construct an explicit model to derive new direct paths between words, we consider the following iterative procedure (Newman, 2010). First consider a walk of a maximum length r where $\mathbf{I}$ is the identity matrix and a "damping parameter" $\alpha < 1$ governs the extent to which new paths are dominated by short paths or by longer paths. During each iteration, indirect links reflecting paths of length r are added to the graphs, producing this sequence of "augmented" graphs:

$$\begin{aligned} \mathbf{G_{rw}}^{(r=0)} &= \mathbf{I} \\ \mathbf{G_{rw}}^{(r=1)} &= \alpha\mathbf{P} + \mathbf{I} \\ \mathbf{G_{rw}}^{(r=2)} &= \alpha^2\mathbf{P}^2 + \alpha\mathbf{P} + \mathbf{I} \end{aligned} \tag{1}$$

In these expressions, longer paths receive lower weights due to operation of the α parameter. The probability of an associative chain surviving across r links is thus α^r. The smaller the value of α, the larger the contribution made by very short paths. This "decay" parameter serves an important role to limit the spread of activation and avoid the entire network to become quickly activated. In the limit, where we consider paths of arbitrarily long length (and accordingly, arbitrarily low weight) we obtain the following expression:

$$\mathbf{G_{rw}} = \textstyle\sum_{r=0}^{\infty}(\alpha\mathbf{P})^r = \mathbf{I} - \alpha\mathbf{P}^{-1} \tag{2}$$

At this point, the "random walk graph" $\mathbf{G}_{\text{rw}}$ combines paths of various lengths obtained from the random walk. However, these paths do not precisely match the associative strength measure proposed earlier, and to address this we apply the exact same procedure that we have used for the other models, namely the PMI$^+$ transformation. Applying 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

To assess how well each of these four models captures human semantic knowledge, we evaluate them using several standard data sets that measure human judgments of similarity and relatedness, and addi-

[2]We did not apply a discount factor for the word association data due to the different characteristics of text corpora and word associations: with smaller data sets the problem of rare words is less pronounced in word associations.

tionally introduce a new data set based on the "remote triad task". We used a variety of different data sets in order to provide insight into *why* some models perform well on some kinds of task and not on others.

4.1 Similarity and relatedness 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-353 similarity data set (Agirre et al., 2009)[3] and the SimLex-999 data (Hill et al., 2016). In the remaining studies people 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 Amazon Mechanical Turk 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).

4.2 Remote triads task

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. (2016). 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 [4].

4.3 Additional details

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 $\alpha = 0.75$, similar to previous studies (De Deyne et al., 2016). 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., in press), and the count model outperformed *word2vec* on the remote triads data.

5.1 Other versions of the I-language models

Given the superiority of the I-language models over E-language models in predicting human responses, it is natural to ask why this occurs. Our word association data arise from a task that elicited multiple judgments from each person. To estimate the effect of the multiple judgment procedure, we restricted the training data to the first associate only (using $\mathbf{G_1}$ rather than $\mathbf{G_{123}}$). After doing so the average

[3] Note that the items were determined by post-hoc raters who split the original WordSim-353 data set in related and similar items. As such, these judgments might not consist of "pure" similarity judgments.

[4] The data are available at `http://simondedeyne.me/data`

Table 1: Spearman rank order correlations between human relatedness and similarity judgments, and the predictions from all four models described earlier. Word association results presented here are based on $\mathbf{G_{123}}$. Further details for $\mathbf{G_1}$ are available in the text.

			Text Corpus		Word Associations	
Data set	*n*	*n(overlap)*	Count	*word2vec*	Count	Random Walk
WordSim-353 Related	252	207	.67	.70	.77	.82
WordSim-353 Similarity	203	175	.74	.79	.84	.87
MTURK-771	771	6788	.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

correlation for the count model fell from .78 to .67, with a much smaller decline (from .82 to .78) for the random walk model. The difference is illuminating: $\mathbf{G_1}$ is much sparser than $\mathbf{G_{123}}$, allowing indirect paths to have a larger impact, which is why the random walk model is more robust than the count model.

A different question to ask is whether our new data set encoded in $\mathbf{G_{123}}$ produces better results than previous ones. There appears to be some modest evidence for this: when using the Edinburgh Association Thesaurus (EAT) consisting of 8,400 words (Kiss et al., 1973), the count model produced an average correlation of .65 to the test data, and the random walk model correlated at .74, both of which are smaller than the values obtained (.78 and .81) using a matched $\mathbf{G_{123}}$ that contains the same words as the EAT. A similar exercise using the USF association data (Nelson et al., 2004) produced correlations of .65 and .77 for the count and random walk model compared to .78 and .82 for the matched $\mathbf{G_{123}}$.

5.2 Other versions of the E-language models

Previous papers have discussed the performance of the E-language models (Levy et al., 2015), but a few additional comments are worth making. Analogous to the effect that the training data have on the I-language models, one might wonder if the poorer performance of *word2vec* was due to the text corpus we used to extract semantic vectors. To test this, we relied on recently published semantic vectors from Mandera et al. (in press)[5], Levy et al. (2015)[6] and Pennington et al. (2014)[7] including items part of $\mathbf{G_{123}}$. For the GloVe vectors based on 6 billion tokens from Pennington et al. (2014), the best result was found for 300 dimensions; the average correlation was .64. Using the GloVe vectors for a 64 billion tokens corpus improved the correlation to .67 and using the GloVe vectors for an enormous corpus of 840 billion tokens to .70. Compared with the published results from Levy et al. (2015), the average correlation was .65. Using the best-fitting vector spaces from Mandera et al. (in press), the correlation was .69 for the published vectors derived from English subtitles using 300 dimensions. Given this, it does not seem likely that the problem was our specific choice of corpus or hyperparameters.

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,

[5] http://zipf.ugent.be/snaut-downloads/spaces/english/predict/

[6] https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/

[7] http://nlp.stanford.edu/projects/glove/

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.

6.1 Cognitive plausibility of E-language models

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., in press). 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 in light of the fact that 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.

6.1.1 Relation between relatedness and similarity

A final remaining question is how we can interpret the lower results for similarity judgments in one of the tasks (SimLex-999) across all models. Does this indicate a fundamental shortcoming in the models?

In this case, the answer depends. Similarity might be important in a NLP setting, for example in constructing thesauri, but the role of similarity in human semantic cognition is mostly an empirical matter. If anything, a variety of studies support a prominent role for relatedness. This includes semantic priming effects when processing a word proceeded by a related one (Hutchison, 2003), event-related potentials in EEG that are triggered by related but not similar words (Kutas and Hillyard, 1984) and fMRI studies that map the structure of the mental lexicon in a more thematic rather than taxonomic way based on similarity (Huth et al., 2016). Add to this that similarity can only be derived for certain concept combinations, and similarity ratings tend to be less reliable than relatedness ratings (Hill et al., 2016), suggests that relatedness judgments have broader use in studies of human semantic cognition.

6.2 Future directions

One obvious way to improve E-language models is by including non-linguistic information as well. As mentioned in the introduction, access to imagery contributes to the responses people give in a word association task and this might explain the fact that I-language models perform very well on the basis of a small number of words. While recent studies have shown some promising results by enhancing language models with visual representations, there's still room for improvement. For example, Bruni et al. (2012) reported findings of a multimodal model that uses word embeddings combined with features extracted from images. An evaluation using the MEN dataset resulted in a correlation of .78 for the best performing model, which is considerably lower than current results for the E-language and especially the I-language models.[8] On the upside, if I-language models do considerably better because they have a privileged access to imagery compared to E-language model, this would also suggest that further improvements by constructing more elaborate multimodal representations are possible. More generally, because the both I-language and E-language models use the same kind of symbolic language-based representations, determining what kind of features (perceptual or other) make the I-language models so successful with very little data might also provide us with valuable pointers towards further refining existing E-language models and NLP applications build from them.

Going forward will also require us to increase the discriminatory power of existing benchmarks, by including judgments that focus on finer perceptual distinctions, or by capitalizing on how humans infer additional structure beyond what's available in E-language. In this study, we have shown that the remote

[8] Unfortunately we did not have access to the semantic vectors from Bruni et al. (2012) so we could not verify whether this is due to the fact that some items were excluded in our experiments.

triad task might be ideally suited to test how well a model can generalize beyond the input and provide a benchmark capable of differentiating competing models. Apart from more sophisticated and more realistic approximating of the E-language environment, there's also a need for better I-language models. While the new word association data addresses some issues from previous studies, future work will aim to include at least 20,000 different cues. Furthermore, improvements might be achieved by looking at native speakers only, applying differential weights to the primary, secondary and tertiary responses, or designing more elaborate spreading activating mechanisms. As such, further gains for association based I-language models would not be a surprise and might help us bridge the external and internal language world.

Acknowledgments

Special thanks to Gert Storms and Marc Brysbaert for supporting the English association data collection. Salary support for this research was provided to Simon De Deyne from ARC grant DE140101749, to Amy Perfors from ARC grant DE120102378, and to Daniel J. Navarro from ARC grant FT110100431.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. 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. Association for Computational Linguistics.

Jean Aitchison. 2012. *Words in the mind: An introduction to the mental lexicon.* Wiley-Blackwell.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. 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.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. 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.

Marc Brysbaert, Emmanuel Keuleers, and Boris New. 2011. Assessing the usefulness of google books word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2.

John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82:407–428.

Mark Davies. 1990–present. The Corpus of Contemporary American English: 520 million words.

Mark Davies. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries.

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

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

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

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

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. 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.

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

Keith A. Hutchison. 2003. Is semantic priming due to association strength or feature overlap? *Psychonomic Bulletin and Review*, 10:785–813.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Leo Katz. 1953. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43.

George Kiss, Christine Armstrong, R. Milroy, and J. Piper. 1973. The computer and literacy studies. chapter An associative thesaurus of English and its computer analysis, pages 153–165. Edinburgh University Press, Edinburgh.

Marta Kutas and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307:161–163.

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

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

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

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. in press. 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.

Ken McRae, Saman Khalkhali, and Mary Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. In *The adolescent brain: Learning, reasoning, and decision making*, pages 39–66. American Psychological Association.

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. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182.

Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. 2007. Asymmetric association measures. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007)*.

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

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers*, 36:402–407.

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

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.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. 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.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

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

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

Lorand B Szalay and James Deese. 1978. *Subjective meaning and culture: An assessment through word associations*. Lawrence Erlbaum Hillsdale, NJ.

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

Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Manfred Wettler, Reinhard Rapp, and Peter Sedlmeier. 2005. Free word associations correspond to contiguities between words in texts*. *Journal of Quantitative Linguistics*, 12(2-3):111–122.