A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data

Francisco Pereira[1], Samuel Gershman[2], Samuel Ritter[3] and Matthew Botvinick[4]

[1]Medical Imaging Technologies, Siemens Healthcare, francisco-pereira@siemens.com
[2]Dept. of Psychology / Center for Brain Science, Harvard University, gershman@fas.harvard.edu
[3]Princeton Neuroscience Institute, Princeton University., swritter@princeton.edu
[4]Google DeepMind, London UK, botvinick@google.com

Abstract

In this paper we carry out an extensive comparison of many off-the-shelf distributed semantic vectors representations of words, for the purpose of making predictions about behavioural results or human annotations of data. In doing this comparison we provide a guide to how vector similarity computations can be used to make such predictions. We also introduce many resources available, both in terms of datasets and of vector representations. Finally, we discuss the shortcomings of this approach and future research directions that might address them.

1 Introduction

1.1 Distributed semantic representations

We are interested in one particular aspect of conceptual representation—the meaning of a word – insofar as it is used in the performance of semantic tasks. The study of concepts in general has a long and complex history, and we we will not attempt to do it justice here (see Margolis & Laurence, 1999; G. L. Murphy, 2002). Researchers have approached the problem of modelling meaning in diverse ways. One approach is to build representations of a concept – a word used in one specific sense – by hand, using some combination of linguistic, ontological and featural knowledge. Examples of this approach include WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), Cyc (Lenat, 1995), and semantic feature norms collected by various research groups (e.g., McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008). An alternative approach, known as *distributional semantics*, starts from the idea that words occurring in similar linguistic contexts – sentences, paragraphs, documents – are semantically similar (see Sahlgren, 2008, for a review). A major practical advantage of distributional semantics is that it enables automatic extraction of semantic representations by analyzing large corpora of text. Since the computational tasks we are trying to solve (and the more general problem of concept representation in the brain) require models that are general enough to encompass the entire English vocabulary as well as arbitrary linguistic combinations, our focus will be on distributional semantic models. Existing hand-engineered systems cannot yet be used to address all the tasks that we consider.

Common to many distributional semantic models is the idea that semantic representations can be conceived as vectors in a metric space, such that proximity in vector space captures a geometric notion of semantic similarity (Turney & Pantel, 2010). This idea has been important both for psychological theorizing (Howard, Shankar, & Jagadisan, 2011; Landauer & Dumais,

1997; Lund & Burgess, 1996a; McNamara, 2011a; Steyvers, Shiffrin, & Nelson, 2004) as well as for building practical natural language processing systems (Collobert & Weston, 2008; Mnih & Hinton, 2007; Turian, Ratinov, & Bengio, 2010). However, vector space models are known to have a number of weaknesses. The psychological structure of similarity appears to disagree with some aspects of the geometry implied by vector space models, as evidenced by asymmetry of similarity judgments and violations of the triangle inequality (Griffiths, Steyvers, & Tenenbaum, 2007; Tversky, 1977). Furthermore, many vector space models do not deal gracefully with polysemy or word ambiguity (but see Jones, Gruenenfelder, & Recchia, 2011; Turney & Pantel, 2010). Recently, a number of different researchers have started focusing on producing vector representations for specific meanings of words (Huang, Socher, Manning, & Ng, 2012; Neelakantan, Shankar, Passos, & McCallum, 2015; Reisinger & Mooney, 2010; Yao & Van Durme, 2011), but these are still of limited use without some degree of manual intervention to pick which meanings to use in generating predictions. We discuss these together with other available approaches in Section 2.1. In the work reported here, we do not attempt to address these issues directly; our goal is to compare the effectiveness of different vector representations of words, rather than comparing them with other kinds of models.

1.2 Modelling human data

Ever since (Landauer & Dumais, 1997) demonstrated that distributed semantic representations could be used to make predictions about human performance in semantic tasks, numerous researchers have used measures of (dis)similarity between word vectors – cosine similarity, euclidean distance, correlation – for that purpose. There are now much larger test datasets than the TOEFL synonym test used in (Landauer & Dumais, 1997), containing hundreds to thousands of judgments on tasks such as word association, analogy, and semantic relatedness and similarity, as described in Section 2.3. The availability of LSA as a web service[1] for calculating similarity between words or documents has also allowed researchers to use it as a means of obtaining a kind of "ground truth" for purposes such as generating stimuli (e.g. (Green, Kraemer, Fugelsang, Gray, & Dunbar, 2010)). In parallel with all this work, researchers within the machine learning community have developed many other distributed semantic representations, mostly used as components of systems carrying out a variety of natural language processing tasks, ranging from information retrieval to sentiment classification (Wang & Manning, 2012).

Beyond behavioural data, distributed semantic representations have been used in cognitive neuroscience, in the study of how semantic information is represented in the brain. More specifically, they have been used as components of forward models of brain activation, as measured with functional magnetic resonance imaging (fMRI), in response to semantic stimuli (e.g. a picture of an object together with the word naming it, or the word alone). Such models learn a mapping between the degree to which a dimension in a distributed semantic representation vector is present and its effect on the overall spatial pattern of brain activation. These models can be inverted to *decode* semantic vectors from patterns of brain activation, which allow validation of the mappings by classifying the mental state in new data; this can be done by comparing the decoded vectors with "true" vectors extracted from a text corpus.

Reviewing this literature is beyond the scope of this paper, but we will highlight particularly relevant work. The seminal publication in this area is (Mitchell et al., 2008), which showed that it was possible to build such forward models, and use them to make predictions about new imaging data. They represented concepts by semantic vectors where dimensions corresponded to different verbs; the vector for a particular concept was derived from co-occurrence counts of the word naming the concept and each of those verbs (e.g. the verb "run" co-occurs more often with animate beings than inanimate objects). Subsequently, (Just, Cherkassky, Aryal, & Mitchell, 2010) produced more elaborate vectors from human judgments, with each dimension

[1] http://lsa.colorado.edu

corresponding to one of tens of possible semantic features. In both cases, this allowed retrospective interpretation of patterns of activation corresponding to each semantic dimension (e.g. ability to manipulate corresponded to activation in motor cortex). Other groups re-analyzing the data from (Mitchell et al., 2008) showed that superior decoding performance could be obtained by using distributed semantic representations rather than human postulated features (e.g. (Pereira, Detre, & Botvinick, 2011) (Liu, Palatucci, & Zhang, 2009)). In particular, (Pereira et al., 2011) used a topic model of a small corpus of Wikipedia articles to learn a semantic representation where each dimension corresponded to an interpretable dimension shared by a number of related semantic categories. Furthermore, the semantic vectors from brain images for related concepts exhibited exhibited similarity structure that echoed the similarity structure present in word association data, and could also be used to generate words pertaining to the mental contents at the time the images were acquired. A systematic comparison of the effectiveness of various kinds of distributed semantic representations in decoding can be found in (B. Murphy, Talukdar, & Mitchell, 2012b). This work has led researchers to consider distributed semantic representations as a core component of forward models of brain activation in semantic tasks, or even try to incorporate brain activation in the process of learning a representation (Fyshe, Talukdar, Murphy, & Mitchell, 2014). The pressing question, from that perspective, is whether representations contain enough information about the various aspects of meaning that might be elicited by thinking about a concept. This question was tackled in (Bullinaria & Levy, 2013), and the authors concluded that the representations currently in use are already very good for decoding purposes, and that the quality of the fMRI data is the main limit of what can be achieved with current approaches. As this conclusion was drawn from datasets containing activation images in response to a few tens of concrete concepts, we believe that we should not look at fMRI to try to gauge the relative information content of different representations; rather, we should use behavioural data to the extent possible, over words naming all kinds of concepts that might be stimuli in experiments. This was the original motivation for this paper.

Our first goal is thus to evaluate how suitable different distributed semantic representations are for reproducing human performance on behavioural experiments or to predict human annotations of data from such tasks. We restrict the comparison to available off-the-shelf representations, because we believe many researchers cannot, or would rather not, go through the trouble of producing their own from a corpus of their choice. As we will see later, the size of the corpus used in producing a representation is a major factor in the quality of the predictions made, and this makes such production logistically complicated, at the very least (because of preparation effort, running time, memory required, etc). In the same spirit, we would like to have our comparison also act as tutorial that shows such predictions can be made and contrasted across representations.

Our second, related goal, is to determine how appropriate vector similarity is for modelling such data, as tasks become more varied and complex. To that effect, we carried out comparative experiments across a range of tasks – word association, relatedness and similarity ratings, synonym and analogy problems – for all the most commonly used off-the-shelf representations.

To do this, we had to assume that the information derived from text corpora suffices to make behavioural predictions; existing literature, and our own experience, tell us that this is the case. But is this the *same* semantic information that would be contained in semantic features elicited from human subjects, for instance? Does it bear any resemblance to the actual representations of stimuli created in the brain while semantic tasks are performed? Can we even say that there is a single representation, or could it be mostly task dependent? Carrying out practical tasks such as sentiment detection using a distributed semantic representation does not require the answer to any of these questions, and neither does decoding from fMRI. The collection of feature norms such as (McRae et al., 2005) has sometimes led to semantic feature representations being viewed as more "real" or "interpretable" than distributed semantic representations. It is possible to constrain the problem of estimating a distributed semantic representation so that the resulting

dimensions look more like semantic features (e.g. values are positive or lie in the [0,1] range, the semantic vectors are sparse), as shown by (B. Murphy, Talukdar, & Mitchell, 2012a). Another issue comes from the fact that representations are derived from some for word co-occurrence, as we shall see later. Co-occurrence of two words in similar contexts does not mean they are equivalent, even though their semantic vectors might be similar (e.g. "happy" and "sad", which would both appear in sentences about emotions or mental states). Hence, some behavioral predictions may not be feasible at all. The question of what information can be captured by semantic features but not distributed semantic representations is discussed at great length in (Riordan & Jones, 2011), where the authors conclude that the amount of information contained in the latter is underestimated. Given that our objective is to compare the ability to generate reasonable predictions from off-the-shelf representations, we will sidestep these questions.

1.3 Related work

The most closely related work is (Baroni, Dinu, & Kruszewski, 2014), an extremely thorough evaluation focusing on many of the same tasks and using many of the same representations, carried out independently from ours. Whereas their main goal was to compare context-counting with context-prediction methods for deriving semantic vectors, our focus is more on helping readers choose from existing representations for use in predictions of behavioural data, as well as showing them how this can be done in practice. To that effect, we have included additional datasets of psychological interest and more vector representations in our comparison. We do recommend that the reader interested in the technical details of the relationships between the different types of method refer to this paper, and also to (Pennington, Socher, & Manning, 2014), (Goldberg & Levy, 2014) and (Levy & Goldberg, 2014). (Griffiths et al., 2007) compares LSA distributed semantic representations with those from a different approach where each word is represented as vector of topic probabilities (within a topic model of a corpus), over word association, the TOEFL synonym test from (Landauer & Dumais, 1997) and a semantic priming dataset. This paper is perhaps the most comprehensive in terms of discussing the suitability of distributed semantic representations for making predictions in psychological tasks, but does not consider most modern off-the-shelf representations or recently available datasets. Finally, (Turney & Pantel, 2010) provides a survey of the uses of distributed semantic representations to carry out tasks requiring semantic information. Both this paper, (Bullinaria & Levy, 2007) and (Rubin, Kievit-Kylar, Willits, & Jones, 2014) cover crucial aspects of processing of text corpora that affect the quality of the representations learned.

2 Methods

2.1 Word representations

The vector space representations we consider are *word* representations, i.e. they assign a vector to a word which might name one or more concepts (e.g. "bank"). They have been chosen *both* because they were public and easily available, and also because they have been used to make predictions about behavioural data or human-generated annotations, or as the input for other procedures such as sentiment classification. We have not included some classic methods, such as HAL (Lund & Burgess, 1996b), COALS (Rohde, Gonnerman, & Plaut, 2006), BEAGLE (Jones, Kintsch, & Mewhort, 2006) or PMI (Recchia & Jones, 2009), primarily because the semantic vectors produced are not publicly available (although the software for producing BEAGLE[2] and PMI[3] models from a given corpus is). These and other methods used specifically to study human performance in semantic tasks are reviewed in detail in (McNamara, 2011b).

[2] https://github.com/mike-lawrence/wikiBEAGLE

[3] http://www.twonewthings.com/lmoss.html

One observation that is often made is that word representations are inherently flawed, in that each vector reflects the use of the corresponding word in multiple contexts with possibly different meaning (see, for instance, Kintsch (2007) for a discussion). It is still possible to use the words in the light of this as, for instance, the vector similarity measure can be driven primarily by values in two vectors present due to related meanings. That said, there have been multiple attempts to solve this problem by producing representations that comprise multiple vectors for each word, corresponding to the different meanings of the word. e.g. (Neelakantan et al., 2015) provides vectors and a description of existing approaches, such as (Reisinger & Mooney, 2010),(Yao & Van Durme, 2011) or (Huang et al., 2012). We have not included these because they would require tagging each stimulus word in the evaluation tasks we consider with the specific meaning present, if available, and this would need to be done separately for each representation type. This is straightforward, though time-consuming, for representations that develop as many vectors as there are senses for a word in WordNet, say. It becomes more complicated when the number of senses is discovered from data, in alternation or simultaneously with the process of generating semantic vectors. In the latter situation, each word-meaning vector must be interpreted in the light of the other word-meaning vectors that it is most similar to (e.g. "apple#1" might be most similar to "computer#2", whereas "apple#2" might be most similar to "orange#1").

We would like to stress that this is a comparison between the semantic vector representations produced by each method operating on a particular corpus, with a given pre-processing and method-specific options. The latter range from the dimensionality of the vectors produced to how information about words and their contexts in the corpus documents is collected and transformed. The choices in all of these factors will affect the performance of the representations in a comparison. Ideally, we would be comparing the representations produced with multiple methods operating on the *same* corpus, and optimizing the various factors for each method. This, however, is a far more demanding endeavour than the current comparison, in terms of both computational resources and time. As we will see later, the best performing representations have all been trained in very large corpora, beyond the scope of what is practical with a single desktop computer. In Section 4 we discuss when it might make sense to learn a representation from scratch, and provide pointers to useful resources covering pre-processing, design choices and trade-offs, toolkits to make the process simpler, as well as the most commonly used corpora.

Across methods, a key distinction is often made between local and global context. When we say that "similar words occur in similar contexts", this typically means one of two things: words are semantically similar if they occur nearby in the same *sentence* (local context) or in the same *document* (global context). There are many variations on this distinction that blur these lines (e.g., contexts that can extend across sentence boundaries, as in N-gram models, or the document considered is a section or a paragraph of a larger ensemble). The models we consider are, for the most part, local context methods, although they use local information in different ways. For more on this distinction, possible variations and impact on the performance representations, please refer to (Bullinaria & Levy, 2007), (Turney & Pantel, 2010) and (Rubin et al., 2014). A second distinction that is made is between *context-counting* and *context-prediction methods*. The former consider co-occurrence counts between words – in whatever context – when producing vectors; the latter use some criterion that reflects the extent to which one can use one word in a context to predict others (or vice versa). This distinction is discussed at great length in (Baroni et al., 2014), as well as (Goldberg & Levy, 2014).

Latent Semantic Analysis (LSA) Latent semantic analysis (LSA; Landauer & Dumais, 1997) is a global context method that models co-occurrence of words within a document. The representation of each word is learned from a $\#words \times \#documents$ word count matrix, containing the number of times a word appeared in each available document context. This matrix is then transformed by replacing each count by its log over the entropy of the word across all contexts. Finally, the matrix undergoes a singular value decomposition; the left singular vectors are word representations in a low-dimensional space, and the right singular vectors are document

representations. The original model was trained on a relatively small corpus of approximately 30000 articles from an encyclopedia. Since we could not obtain that model, we use a random sample of 10% of the articles from Wikipedia (aproximately 300000 articles) to generate vector representations with same number of dimensions (300).

Multi-task neural network embedding (CW) In the article introducing this method (Collobert & Weston, 2008) the authors introduce a convolutional neural network architecture to carry out a series of language processing tasks – e.g. part-of-speech tagging, named entity tagging, semantic role labeling – from a vector representation of the words in a sentence. The vector representation was learned using a local context approach, by training a model that used it to assign higher probability to the right word in the middle of the window than to a random one, making this an early instance of "predict" model using a window of 5 words in each direction. The model was trained on a large subset of Wikipedia containing approximately 631 million words, and distributed as 25-, 50-, 100- and 200-dimensional representations. The vectors were obtained from a third-party web site[4].

Hierarchical log-bilinear model (HLBL) This and other related local context methods were introduced in (Mnih & Hinton, 2007) and further developed in (Mnih & Hinton, 2008). The commonality between them is learning vector representations for use in a statistical model for predicting the conditional distribution of a word given a window of 5 preceding ones. The model was trained on a subset of the Associated Press dataset containing approximately 16 million words. The representations available are 100- and 200-dimensional. The vectors were obtained from the same site as those in representation CW.

Non-negative sparse embedding (NNSE) This approach was introduced in (B. Murphy et al., 2012a) and aims at learning a word representation that is sparse (vector values are positive, and zero for most dimensions) and disjoint (positive vector values tend to be disjoint between word types such as abstract nouns, verbs and function words). NNSE is a global context model, in that it models co-occurrence of words within a document (similarly to (Landauer & Dumais, 1997)) but it combines these with word-dependency co-occurrence counts (similarly to (Lund & Burgess, 1996b)). The counts were normalized by a transformation into positive pointwise mutual information (positive PMI) scores (Bullinaria & Levy, 2007; Turney & Pantel, 2010). The process of generating the vectors with the desired properties is similar to that of (Landauer & Dumais, 1997), albeit with a more complex process of factorization of word-by-score matrices, which is beyond the scope of this paper. It was learned from a subset of the Clueweb dataset (approximately 10 million documents and 15 billion words). The representations available are 50-, 300-, 1000- and 2500-dimensional and the vectors are provided by the authors[5].

Word2vec (W2VN) The representations in this class are produced with local context methods, trained on the Google News dataset (approximately 100 billion words) and distributed by Google[6]. The continuous bag-of-words model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) (a.k.a. CBOW or "negative sampling") learns to predict a word based on its context (predict the word in the middle of a context window based on an average of the vector representations of the other words in the window. The representation derived with this model is 300-dimensional. The continuous skipgram model (Mikolov, Chen, Corrado, & Dean, 2013) learns to predict words in the context window from a word in the middle of it. In the publicly available distribution produced with the latter method, the vectors distributed do not correspond to individual words, but rather to identifiers for entities from the Freebase ontology for a particular meaning of each word. Hence, we excluded this variant of the word2vec method from the comparison. A good explanation of the two methods is provided in (Goldberg & Levy, 2014), and (Levy & Goldberg, 2014) derives a connection between the skipgram approach and factorization of PMI count matrices.

[4] metaoptimize.com/projects/wordreprs

[5] http://www.cs.cmu.edu/~bmurphy/NNSE/

[6] https://code.google.com/p/word2vec/

Global Vectors (GV300, GV42B, GV840B) This approach is described in (Pennington et al., 2014). It is a global context method, in that it models co-occurrence of words across a corpus, in a manner similar to (Burgess, 1998; Rohde, Gonnerman, & Plaut, 2009), but it operates by factorizing a transformed version of the term-term co-occurrence matrix. The factorization is similar to that used in (Landauer & Dumais, 1997) but the transformation is beyond the scope of this paper; performance asymptotes with co-occurrences windows of 8-10 words. The model is trained on a combination of Wikipedia 2014 and Gigaword 5 (6 billion tokens) or Common Crawl (42 billion and 840 billion tokens). All three versions are 300-dimensional and made available by the authors[7].

Context-counting (BDKC) and context-predicting (BDKP) vectors This approach is described in (Baroni et al., 2014). The vector representations were extracted from a corpus of 2.8 billion tokens, constructed by concatenating ukWAC, the English Wikipedia and the British National Corpus, using two different methods ("count" and "predict"). Both methods use local context windows, although in different ways. The "count" models were obtained by decomposing matrices of word co-occurrence counts within a window of size 5 using SVD, after transforming word count scores to PMI.. The "predict" models were obtained by using the word2vec software on the same corpus, using the CBOW approach with a window of size 5. In both cases the authors used their own toolbox, DISSECT[8], to produce the vector representations. For our comparison we use the best "predict" and "reduced count" representations made available by the authors[9], which are 400- and 500-dimensional, respectively.

2.2 Vector similarity measures

Our underlying hypothesis is that vector similarity in the space used to represent concepts reflects semantic relatedness. We consider three kinds of measures – euclidean distance, correlation and cosine similarity – and we will use the term "similarity" to mean either high similarity proper or low distance, depending on the measure used.

The euclidean distance between n-element vectors $\mathbf{u}$ and $\mathbf{v}$ is

$$\|\mathbf{u}-\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^{n}(u_i-v_i)^2}.$$

The correlation similary between vectors $\mathbf{u}$ and $\mathbf{v}$ is

$$\text{correlation}(\mathbf{u},\mathbf{v}) = \frac{\sum_{i=1}^{n}(u_i-\mu_u)(v_i-\mu_v)}{\sigma_u\sigma_v} = \mathbf{u}^*\mathbf{v}^*$$

where μ_u and σ_u are the mean and standard deviation of vector $\mathbf{u}$, respectively (and analogously for vector $\mathbf{v}$). If we consider the normalization where each vector is z-scored, i.e. $\mathbf{u}^* = \frac{\mathbf{u}-\mu_u}{\sigma_u}$, the correlation can be viewed as a product of normalized vectors.

The cosine similarity between vectors $\mathbf{u}$ and $\mathbf{v}$ is

$$\text{cosine}(\mathbf{u},\mathbf{v}) = \frac{\sum_{i=1}^{n} u_i v_i}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} = \mathbf{u}'\mathbf{v}'.$$

where $\|\mathbf{u}\|_2$ is the length of vector $\mathbf{u}$ (and analogously for vector $\mathbf{v}$). If we consider the normalization where each vector is made length 1, i.e. $\mathbf{u}' = \frac{\mathbf{u}}{\|\mathbf{u}\|_2}$, the cosine similarity can be viewed as a product of normalized vectors.

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

[8] http://clic.cimec.unitn.it/composes/toolkit

[9] http://clic.cimec.unitn.it/composes/semantic-vectors.html

Given that correlation and cosine similarity are, in essence, vector products, operating on implicitly normalized versions of the original vectors, they are invariant – to a degree – to the magnitude of the vector entries or the length of the vector. This is not the case for euclidean distance, and one of our goals is to determine whether this is a relevant factor for the applications we have in mind. Furthermore, values of euclidean distance between vectors are not directly comparable across different vector representations. The use of evaluation tasks based on rankings, where the score is the position of a "correct" answer, is meant to allow a single approach that works regardless of the similarity measure used.

The use of rank measures allows us to avoid having to directly predict the raw scores obtained from judgements of annotations. However, it could also mask large differences in the relationship between scores for different items across different models (e.g. a representation where vectors for small sets of words are very similar to each other and dissimilar to all else, versus one with more graded similarity values between the same words). An alternative approach would be to use a technique such as the Luce choice rule, which can be used to convert both the scores and the distances/similarities produced from any representation into the same normalized scale. This approach and potential pitfalls of using vector distance/similarity are discussed in (Jones et al., 2011) (and, as described earlier, in (Griffiths et al., 2007)).

2.3 Datasets used in evaluation tasks

The data are available online and pointers to the original paper and a brief description are provided in each section.

2.3.1 Word association

Nelson, McEvoy, and Schreiber (2004) collected free association norms for 5000 words.[10] Over 6000 participants were asked to write the first word that came to mind that was meaningfully related or strongly associated to the presented word. The word association data were then aggregated into a matrix form, where S_{ij} represents the probability that word j is the first associate of word i. The dataset is distributed in a reduced dimensionality version containing 400-dimensional vectors for each word, from which we re-assembled the entire association matrix. Our hypothesis, following the work of Steyvers et al. (2004), is that word association can be predicted by vector similarity. Prediction accuracy is measured as the proportion of the top 1% associates for a particular word that are also in the top 1% of words ranked closest in vector space to that word, averaged over all words. This criterion is different from but related to the one in (Griffiths et al., 2007), where the authors considered the probability of the *first* associate being present when considering the top m words, for varying values of m; hence the results are not directly comparably for LSA300, and neither would they be on the grounds of our having used a different corpus (and settings) to learn the representation. We chose our criterion to allow comparison across representations of multiple dimensionalities, and because the top 1% of associates contains most of the probability mass for almost all concepts.

2.3.2 Similarity and relatedness judgments

MEN The MEN dataset[11] consists of human similarity judgments for 3000 word pairs, randomly selected from words that occur at least 700 times in a large corpus of English text, and at least 50 times (as tags) in a subset of the ESP game dataset. It was collected and made public by the University of Trento for testing algorithms implementing semantic similarity and relatedness measures, and introduced in (Bruni, Tran, & Baroni, 2014). Each pair was randomly

[10] http://psiexp.ss.uci.edu/research/software.htm

[11] http://clic.cimec.unitn.it/~elia.bruni/MEN

matched with a comparison pair, and participants were asked to rate whether the target pair was more or less related than the comparison pair. Each pair was rated against 50 comparison pairs, producing a final score on a 50-point scale. Participants were requested to be native English speakers.

SimLex-999 The SimLex-999 dataset[12] was collected to specifically measure similarity, rather than just relatedness or association, of words (e.g. "coast" and "shore" are similar, whereas "clothes" and "closet" are not; both pairs are related). It contains a selection of adjective, verb and noun words pairs, with varying degrees of concreteness. The similarity ratings were produced by 500 native English speakers, recruited via Amazon Mechanical Turk, on a scale going from 0 to 6. The methodology and motivation are further described in (Hill, Reichart, & Korhonen, 2014).

WordSim-353 The WordSimilarity-353 dataset[13] was introduced in (Finkelstein et al., 2001). It contains a set of 353 noun pairs representing various degrees of similarity. 16 near-native English speaking subjects were instructed to estimate the relatedness of the words in each pair on a scale from 0 (totally unrelated words) to 10 (very much related or identical words).

2.3.3 Synonyms and analogy problems

TOEFL This dataset was originally described in (Landauer & Dumais, 1997), and consists of 80 retired items from the synonym portion of the Test of English as a Foreign Language (TOEFL), produced by the Educational Testing Service . Each item consists of a probe word and four candidate synonym words; the subject then picks the candidate whose meaning is most similar to that of the probe. The test items are available upon request from the authors.[14]

Google analogy This dataset was originally introduced in (Mikolov, Chen, et al., 2013) and described in more detail in (Mnih & Kavukcuoglu, 2013) . It contains several sets of analogy problems, of the form "A is to B as C is to ?". They are divided into semantic problems (where A and B are semantically related) and syntactic problems (where A and B are in a grammatical relation). The five semantic analogies come from various topical domains, e.g. cities and countries, currencies and family; the nine syntactic analogies use adjective-to-adverb formation, opposites, comparatives, superlatives, tense and pluralization.

3 Experiments and results

3.1 Prediction of behavioural data or human annotations

Each of the evaluation tasks consists of generating a prediction from words or pairs of words, and their corresponding vectors, which is matched to behavioural or human-annotated data in a task-dependent way described in the rest of this section. The datasets used are described in Section 2.3. The vector representations used are those introduced in Section 2.1, Latent Semantic Analysis (LSA), Multi-task neural network embedding (CW), Hierarchical log-bilinear model (HLBL), Non-negative sparse embedding (NNSE), Global Vectors for word representation (GV), Word2vec (W2VS) and context-counting (BDKC) and context-predicting (BDKP); in each graph, the method abbreviations are followed by a number indicating dimensionality. If a particular word is not in the vocabulary of a representation, the corresponding fragments of data are ignored in tests (this happens for a tiny fraction of each representation, if at all). The results on all tasks are shown in Figure 1, with all performance scores in the range $[0, 1]$ (1 best); the specific performance measure is task-dependent, as described below. Given that results are very similar for cosine and correlation vector similarities, we omit the latter.

[12] http://www.cl.cam.ac.uk/~fh295/simlex.html

[13] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353

[14] http://lsa.colorado.edu

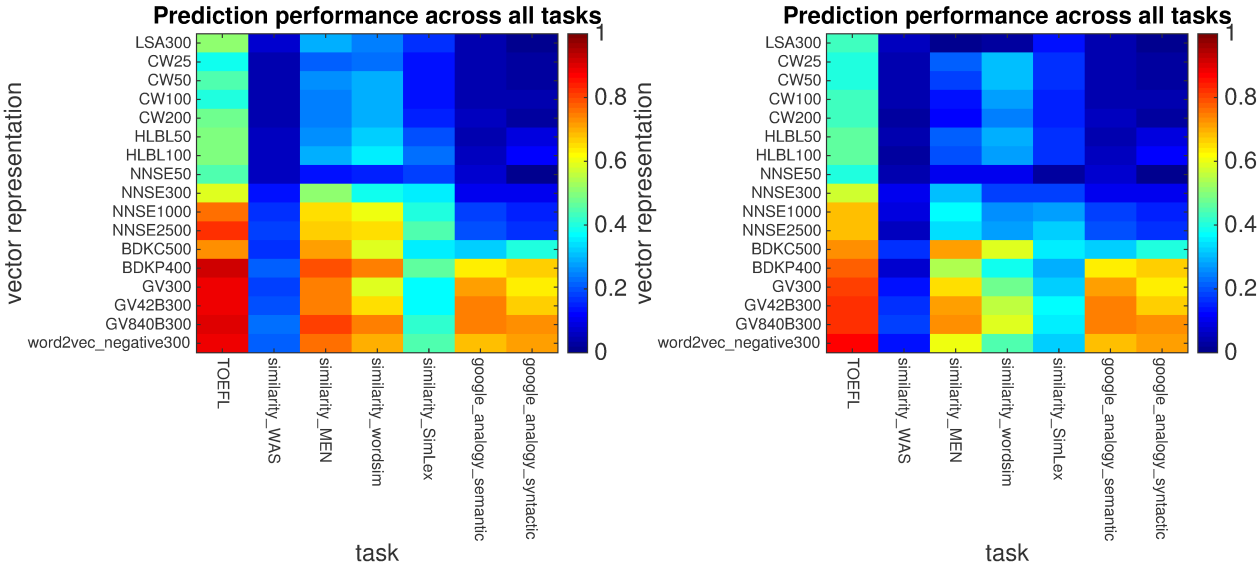

Figure 1: Performance of predictions generated from various vector representations across all tasks available, using cosine similarity (left) or euclidean distance (right). The performance measure is task dependent, but always in the range $[0, 1]$ (1 best). For Google Analogy only cosine results were obtained, because of computational constraints.

3.1.1 Word association

For each word in the dataset, we rank all others by the similarity of their vectors to its vector. The score is the overlap between the top 50 associates of the word and the top 50 most similar vectors. This number was chosen because it is approximately 1% of the total number of words for which data is available; the exact number depends on which words are in the vocabulary associated with a particular vector representation. We chose to consider overlap fraction over the top words, as opposed to a measure like rank correlation, because most of the ranking is of little interest to us (the vast majority of words are barely used as associates, if at all).

3.1.2 Similarity and relatedness judgments

These tasks rely on similarity and relatedness judgments between pairs of words. All tasks share the same evaluation procedure:

1. Rank all test word pairs from most to least similar judgement value ("true ranking")
2. For each test word pair, compute the similarity between their respective vectors, and produce a second ranking from most to least similar ("predicted ranking")
3. The prediction "accuracy" is measured by Spearman rank correlation between the true and predicted rankings positions of all word pairs

The use of Spearman rank correlation allows the results to be comparable across representations and tasks, irrespective of dimensionality or vector similarity measure.

3.1.3 Synonyms and analogy problems

TOEFL For each of the 80 test items, we calculated the similarity between the vector for the probe word and the vectors for the four candidate synonyms, picking the most similar. The

Tasks	*TOEFL*	*WAS*	*MEN*	*wordsim*	*simlex*	*GAsemantic*	*GAsyntactic*
TOEFL	–	0.98	0.97	0.96	0.94	0.91	0.90
WAS	–	–	0.99	0.96	0.96	0.88	0.88
MEN	–	–	–	0.97	0.95	0.87	0.88
wordsim	–	–	–	–	0.94	0.84	0.86
simlex	–	–	–	–	–	0.75	0.76
GAsemantic	–	–	–	–	–	–	0.99
GAsyntactic	–	–	–	–	–	–	–

Table 1: Correlation between the scores for all representations in each pair of tasks.

Tasks	*TOEFL*	*WAS*	*MEN*	*wordsim*	*simlex*	*GAsemantic*	*GAsyntactic*
log_{10} corpus size	0.68	0.75	0.68	0.61	0.66	0.67	0.61
dimensionality	0.42	0.41	0.41	0.46	0.55	0.06	0.03

Table 2: Correlation between representation performance and characteristics (log_{10} of the corpus size (row 1), dimensionality (row 2), across all tasks considered.

prediction accuracy is the fraction of test items for which the correct word was selected.

Google analogy The task is carried out by creating a composite of the vectors for the words in the problem $(A - B + C)$ and then finding the vocabulary word with the most similar vector (using cosine similarity, excluding the vectors for A, B and C). The prediction accuracy is the fraction of test items for which the correct word was selected.

3.2 Experiments aggregating results on individual tasks

Given all the experiments described above, are there broad trends across the results? The first area we considered was the performance of representations across tasks. We quantify the similarity in performance by computing, for each pair of tasks, the correlation between the scores for all representations in each of them. These results are shown in Table 1, and suggest that the relative performance of the models is very similar across all the tasks.

As discussed earlier, we are comparing representations obtained by applying a given method to a given corpus. Our intent is not to compare the methods in isolation, which would require applying them to a benchmark corpus and using the resulting representations. It is, however, still possible to ask whether certain properties of the method or the corpus affect performance in general, *across* representations. To that effect, we carried out two separate experiments. In the first, we correlated the log_{10} of the corpus size for each representation with the result of using it in each task. In the second, we did the same for the dimensionality of the representation. The results are shown in the first and second row of Table 2. Increases in corpus size do appear to lead to consistently better performance. Our conjecture is that this is to exposure to both more instances of each context where two words could conceivably co-occur, and also a greater variation of types of context; these large corpora combine many types of text found on the web – from blogs to news or transcripts – as well as books and encyclopaedia articles. It is less clear that increases in dimensionality are advantageous beyond a certain point. NNSE and "count" vectors do not perform better than other representations with 300 dimensions, and there are underperforming representations with that dimensionality as well.

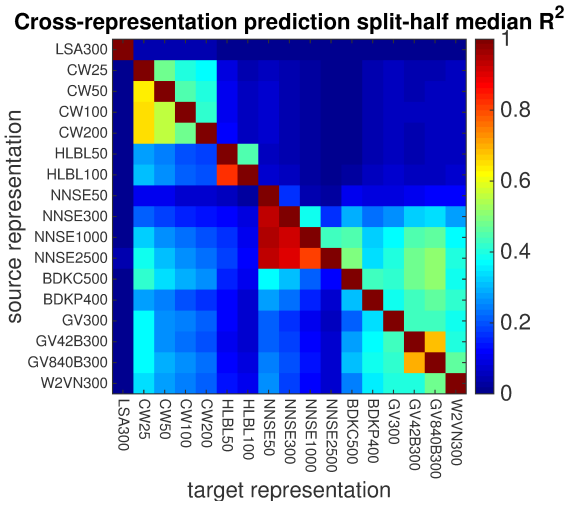

Figure 2: The median R^2 across all dimensions for a regression model predicting each representation (target) from each of the others (source). The R^2 is obtained in split-half cross-validation.

3.3 Relation between representations

As seen earlier, the performance of representations is strongly related to the size of the corpus they were learned from. With that in mind, the best performing representations have comparable performance across tasks, so the question arises of whether they are, in fact, redundant and contain similar information. In practice, this would mean that it would be possible to express each dimension of one representation in terms of the dimensions of another one. We have implemented a simple version of this approach by training ridge regression models to predict each dimension in one representation (target) as a linear combination of the dimensions of another (source). The predictability of the target representation from the source one can then be summarized as the median R^2 across all its dimensions.

The predictions are obtained with an even-odd word split-half cross-validation, in order to remove the confound of having a better result by having more dimensions in the source representation. Furthermore, in order to ensure that the vocabularies used are comparable across representations, and the results pertain primarily to the basic vocabulary used in our evaluation tasks, we restricted the words considered in two ways. The first is that they come from a list of 40K lemmas published by (Brysbaert, Warriner, & Kuperman, 2014); around 30K of these are words or two word compounds that correspond to a lemma in WordNet. Beyond that, we used the words in common across the vocabularies each source and target representation (this ranged from 15K for the smallest representations to close to 30K for those with the largest vocabularies). The results shown in in Figure 2 use $\lambda = 1$ in the ridge regression models, but results are relatively similar for other values of λ a few orders of magnitude above or below.

As expected, median R^2 is high within each representation type, and higher when using the representations with more dimensions to predict those with fewer. NNSE is harder to predict from other representations, because the positivity and sparsity constraints are not enforced in our regression model; therefore, results should not be interpreted as meaning that NNSE has information that other methods cannot predict. Across the representations trained on very large corpora, the median R^2 appears to converge around 0.5. This suggests that there is a residual inability to predict, and possibly nonlinear relationships between dimensions.

4 Discussion

In light of the results presented in the previous section, we conclude that there are representations which perform better across all of our evaluation tasks, namely GloVe and word2vec (and context-predicting vectors, which uses the same method as word2vec, on a different corpus). Given our reliance on measures of vector similarity to generate predictions in each task, it appears correlation and cosine similarity are somewhat better than euclidean distance for this purpose. NNSE at higher dimensionality also performs well across word association and similarity/relatedness tasks, but less so in analogy tasks. Given that this is the only task that requires identifying a correct analogy answer out of a range of 10K-100K possibilities, and the fact that vectors are positive and sparse, it is possible that the vector similarity measures we use are not the most appropriate to allow fine distinctions between closely related words.

From the practical standpoint of a researcher in need of an off-the-shelf word representation, we would thus recommend using word2vec (in its context-predicting vector version) or GloVe. This is because their performance is roughly equivalent and it is straightforward to obtain vectors in a convenient format (word2vec itself requires extracting them from a binary distribution). Both options come with a reasonably sized vocabulary; this matters because too large a vocabulary will lead to intense use of memory for no particular gain, as the words of interest in psychological studies tend to be relatively frequent. Both representations are good for nearest neighbor computations, and simple vector differences capture the meaning of combinations of two words (as suggested by the analogy results). GloVe, in addition, has mostly uncorrelated dimensions. This makes it especially suitable for building other prediction models that work with vectors as inputs, such as regression models of subject ratings. Both representations have very dense vectors. If the target application requires some degree of interpretability, e.g. by identifying which dimensions have the most impact in the prediction, or treating each dimension as a "magnitude" score, it may make more sense to use a representation like NNSE. The vectors are sparse and dimensions are positive and shared by a relatively small set of words.

These conclusions are largely consistent with those of (Baroni et al., 2014), who found that prediction-based, local context methods outperformed co-occurrence count-based methods. The authors compared both classes of approaches over the *same* corpus, systematically varying parameters such as the transformation of the count data or the width of context windows. Although GloVe was not part of that comparison, we do include both the best performing count and predict models from that study (BDKC500 and BDKP400) and a variant of GloVe (GV300) obtained from a corpus of comparable size. From this particular contrast, and given that BDKP400 was the best model in that study, we believe that word2vec-related models may have a slight edge in performance relative to GloVe. This is compensated by increasing the GloVe corpus size; since the large corpus versions are now available off-the-shelf, it's not clear that there is any advantage in choosing one over the other. . Interestingly, both CW and HLBL are local context, prediction-based methods, and perform rather poorly despite being in widespread use. This suggests that the specific local context prediction method and optimization objective do matter, as does the size of the corpus used in training. Given the robust correlation between this and performance of a representation across tasks, it is likely a key factor if considering a representation for "general purpose" use. Increases in dimensionality do not appear to have the same effect on performance. Ultimately, all the methods considered use word co-ocurrence statistics, in what may be a more or less direct fashion. (Pennington et al., 2014) helpfully provide a perspective that connects their model to others, in particular the skipgram version of word2vec introduced in (Mnih & Kavukcuoglu, 2013); a detailed explanation of the two word2vec versions is given in (Goldberg & Levy, 2014).

As mentioned earlier, this is a comparison of off-the-shelf representations, obtained as the result of applying a specific method to a specific corpus. Even though certain representations are superior across all the tasks considered, this does not mean that the methods used to produce

them are necessarily superior. However, the methods that performed best are easily deployable, with well-documented code that allows tuning of parameters such as the context considered for co-occurrence. On that count alone, they are likely to be the best options available.

A separate question is that of when would a researcher want to depart from using an off-the-shelf representation. One situation would be the need for a specialized corpus, from a given technical (e.g. biomedical texts) or otherwise restricted domain (e.g. educational texts for particular grades). Across methods, it would still be the case that words appearing in similar contexts would have similar representations. Given the restricted corpus size, and increased influence of every word co-occurrence, it would be even more important to define context appropriately (e.g. same document might make sense for information retrieval applications, but same passage or sentence would likely make more sense for applications where one wants to represent an individual word). Absent a rationale for picking a specific corpus, the ones most commonly used are Wikipedia[15], ClueWeb12[16], Common Crawl[17] and Gigaword[18] (not freely available, unlike the others). We would recommend starting with the Wikipedia corpus, as there are many tools specifically designed to pre-process or subset it, in a variety of programming languages; furthermore, the case may be made that Wikipedia is especially able to provide contexts covering a "cognitive space" shared by all humans (Olney, Dale, & DMello, 2012).In this situation, it would be worthwhile to consider using gensim[19], as it provides support for some of the corpus processing required and implementations of not just word2vec but also its extensions for representing small text passages (Le & Mikolov, 2014); the DISSECT[20] toolkit provides much overlapping functionality, and may be preferrable to operations relying directly on (transformed) co-occurrence counts (such as SVD or non-negative matrix factorization). The reader interested in doing this should also consult (Bullinaria & Levy, 2007), (Turney & Pantel, 2010), (Bullinaria & Levy, 2012), and (Rubin et al., 2014) for a discussion of the types of context, transformations of counts and several other pre-processing steps and factors that can affect performance of representations.

Overall, the results confirm that vector similarity allows us to make reasonable predictions, and that certain representations are better for this purpose across all tasks considered. More specifically, though, results are good for semantic relatedness, less good for controlled semantic similarity, and even less so for word association. The question is, then, whether this progression reflects a corresponding psychological process increasingly different from something like a nearest neighbour operation in semantic space; many other examples of such issues may be found in (Griffiths et al., 2007; Turney & Pantel, 2010). More recently,(Recchia & Jones, 2009) showed that the use of an appropriate choice rule, when combined with vector space operations, could make it possible for such models to overcome at least some of those issues.

Our goal, however, is not to provide a state-of-the-art approach to making each type of prediction. Our main intention was to see whether certain off-the-shelf representations were preferrable, across a wide range of tasks and using a simple, robust prediction approach that yielded results comparable across them. In doing so, we realized this work could also serve as an introduction to this area, and provide a guide to the publicly available resources used for our evaluation tasks. We can, however, still consider the question of what can or should be predicted. For the purposes discussed in the introduction, such as modelling "generic" semantic activation in brain imaging experiments, these representations provide a clear advantage over human-generated semantic features (Pereira et al., 2011). Further improvement there will likely come from moving to concept rather than word vectors and representing sentences or passages using methods such as paragraph vector (Le & Mikolov, 2014) or skip-thought (Kiros et al.,

[15] https://dumps.wikimedia.org/enwiki

[16] http://www.lemurproject.org/clueweb12.php

[17] https://commoncrawl.org

[18] http://catalog.ldc.upenn.edu/LDC2003T05

[19] https://radimrehurek.com/gensim

[20] http://clic.cimec.unitn.it/composes/toolkit/index.html

2015). For modelling of human performance in behavioural tasks, or annotations, the picture is more complicated. First, it is not even clear that there would be a single representation at work across all tasks. Even if this were the case, it is still possible that task or context of use would modulate the use (e.g. by attending to different semantic features of stimuli, one might correspondingly use different dimensions of a semantic space, or a different similarity function). In the light of this, we hypothesize that further progress will come from modelling what happens in the process of making a judgment. Semantic vectors can still be at the root of this, e.g. as inputs to a model that predicts probability of choosing an associate for a probe word, or which sense of the probe word to use. Other practical issues that are generally ignored – such as instructions given to subjects changing how they produce a judgment – may still allow for the use of vector similarity. One possible approach here would be to use metric learning (Kulis, 2012; Yang & Jin, 2006). More precisely, this would entail weighting each dimension differently when computing a vector similarity, rather than all dimensions equally, essentially changing how vector similarity operates. The weights for the metric would be learned on data from a number of "training" subjects given each of the possible sets of instructions. The learned metrics could then be used to generate predictions on left-out "test" subjects and, if successful, their respective dimension weights analyzed to understand which semantic dimensions played a role in each prediction (and which words, in turn, had loadings in those dimensions).

Acknowledgments

We are grateful to the editor and reviewers for their feedback. In particular, we would like to thank Jon Willits for his very insightful comments regarding almost every part of this article, and his helpful suggestions for additional experiments we included after revision. We are also grateful to Walid Bendris for help in implementing an early version of this evaluation.

Funding

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL), under contract FA8650-14-C-7358. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (Vol. 1, pp. 238–247).

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, *49*, 1–47.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3), 904–911.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510–526.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, *44*(3), 890–907.

Bullinaria, J. A., & Levy, J. P. (2013). Limiting factors for mapping corpus-based semantic representations to brain activity.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the hal model. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 188–198.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on world wide web* (pp. 406–414).

Fyshe, A., Talukdar, P. P., Murphy, B., & Mitchell, T. M. (2014). Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proc. of acl*.

Goldberg, Y., & Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*(1), 70–76.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Hill, F., Reichart, R., & Korhonen, A. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. (2011). Constructing semantic representations from a gradually changing representation of temporal context. *Topics in Cognitive Science*, *3*, 48–73.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1* (pp. 873–882).

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2011). In defense of spatial models of lexical semantics. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3444–3449).

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of memory and language*, *55*(4), 534–552.

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, *5*(1), e8622.

Kintsch, W. (2007). Meaning in context. *Handbook of latent semantic analysis*, 89–105.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3276–3284).

Kulis, B. (2012). Metric learning: A survey. *Foundations and Trends in Machine Learning*, *5*(4), 287–364.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications*

of the ACM, *38*, 33–38.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185).

Liu, H., Palatucci, M., & Zhang, J. (2009). Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th annual international conference on machine learning* (pp. 649–656).

Lund, K., & Burgess, C. (1996a). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208.

Lund, K., & Burgess, C. (1996b). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Margolis, E., & Laurence, S. (1999). *Concepts: Core readings*. MIT Press.

McNamara, D. S. (2011a). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*, *3*, 3–17.

McNamara, D. S. (2011b). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*, *3*(1), 3–17.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, *3*, 235–244.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–5. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/18511683` doi: 10.1126/science.1152876

Mnih, A., & Hinton, G. (2007). Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on machine learning* (pp. 641–648).

Mnih, A., & Hinton, G. E. (2008). A scalable hierarchical distributed language model. In *Nips* (pp. 1081–1088).

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems* (pp. 2265–2273).

Murphy, B., Talukdar, P., & Mitchell, T. (2012b). Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the first joint conference on lexical and computational semantics* (pp. 114–123).

Murphy, B., Talukdar, P. P., & Mitchell, T. (2012a). Learning effective and interpretable semantic models using non-negative sparse embedding. In *Coling* (pp. 1933–1950).

Murphy, G. L. (2002). *The big book of concepts*. MIT press.

Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. (2015). Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 402–407.

Olney, A. M., Dale, R., & DMello, S. K. (2012). The world within wikipedia: an ecology of mind. *Information*, *3*(2), 229–255.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word represen-

tation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Pereira, F., Detre, G., & Botvinick, M. (2011). Generating text from functional brain images. *Frontiers in human neuroscience*, *5*, 72.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, *41*(3), 647–656.

Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Annual conference of the association for computational linguistics* (pp. 109–117).

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, *8*, 627–633.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2009). An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*, 1–33.

Rubin, T. N., Kievit-Kylar, B., Willits, J. A., & Jones, M. N. (2014). Organizing the space and behavior of semantic models. In *Annual conference of the cognitive science society* (Vol. 2014, p. 1329).

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*, 33–54.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, 237–249.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384–394).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*, 183–190.

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (pp. 90–94).

Yang, L., & Jin, R. (2006). Distance metric learning: A comprehensive survey. *Michigan State Universiy*, *2*.

Yao, X., & Van Durme, B. (2011). Nonparametric bayesian word sense induction. In *Proceedings of textgraphs-6: Graph-based methods for natural language processing* (pp. 10–14).