A corpus-based evaluation method for Distributional Semantic Models

Abdellah Fourtassi[1,2]
abdellah.fourtassi@gmail.com

Emmanuel Dupoux[2,3]
emmanuel.dupoux@gmail.com

[1]Institut d'Etudes Cognitives, Ecole Normale Superieure, Paris
[2]Laboratoire de Sciences Cognitives et Psycholinguistique, CNRS, Paris
[3]Ecole des Hautes Etudes en Sciences Sociales, Paris

Abstract

Evaluation methods for Distributional Semantic Models typically rely on behaviorally derived gold standards. These methods are difficult to deploy in languages with scarce linguistic/behavioral resources. We introduce a corpus-based measure that evaluates the stability of the lexical semantic similarity space using a pseudo-synonym same-different detection task and no external resources. We show that it enables to predict two behavior-based measures across a range of parameters in a Latent Semantic Analysis model.

1 Introduction

Distributional Semantic Models (DSM) can be traced back to the hypothesis proposed by Harris (1954) whereby the meaning of a word can be inferred from its context. Several implementations of Harris's hypothesis have been proposed in the last two decades (see Turney and Pantel (2010) for a review), but comparatively little has been done to develop reliable evaluation tools for these implementations. Models evaluation is however an issue of crucial importance for practical applications, i.e., when trying to optimally set the model's parameters for a given task, and for theoretical reasons, i.e., when using such models to approximate semantic knowledge.

Some evaluation techniques involve assigning probabilities to different models given the observed corpus and applying maximum likelihood estimation (Lewandowsky and Farrell, 2011). However, computational complexity may prevent the application of such techniques, besides these probabilities may not be the best predictor for the model performance on a specific task (Blei, 2012). Other commonly used methods evaluate DSMs by comparing their semantic representation to a behaviorally derived gold standard. Some standards are derived from the TOEFL synonym test (Landauer and Dumais, 1997), or the Nelson word associations norms (Nelson et al., 1998). Others use results from semantic priming experiments (Hutchison et al., 2008) or lexical substitutions errors (Andrews et al., 2009). Baroni and Lenci (2011) set up a more refined gold standard for English specifying different kinds of semantic relationship based on dictionary resources (like WordNet and ConceptNet).

These behavior-based evaluation methods are all resource intensive, requiring either linguistic expertise or human-generated data. Such methods might not always be available, especially in languages with fewer resources than English. In this situation, researchers usually select a small set of high-frequency target words and examine their nearest neighbors (the most similar to the target) using their own intuition. This is used in particular to set the model parameters. However, this rather informal method represents a "cherry picking" risk (Kievit-Kylar and Jones, 2012), besides it is only possible for languages that the researcher speaks.

Here we introduce a method that aims at providing a rapid and quantitative evaluation for DSMs using an internal gold standard and requiring no external resources. It is based on a simple same-different task which detects pseudo-synonyms randomly introduced in the corpus. We claim that this measure evaluates the intrinsic ability of the model to capture lexical semantic similarity. We validate it against two behavior-based evaluations (Free association norms and the TOEFL synonym test) on semantic representations extracted from a Wikipedia corpus using one of the most commonly used distributional semantic models : the Latent Semantic Analysis (LSA, Landauer and Dumais (1997)).

In this model, we construct a word-document matrix. Each word is represented by a row, and

Proceedings of the ACL Student Research Workshop, pages 165–171,
Sofia, Bulgaria, August 4-9 2013.

each document is represented by a column. Each matrix cell indicates the occurrence frequency of a given word in a given context. Singular value decomposition (a kind of matrix factorization) is used to extract a reduced representation by truncating the matrix to a certain size (which we call the semantic dimension of the model). The cosine of the angle between vectors of the resulting space is used to measure the semantic similarity between words. Two words end up with similar vectors if they co-occur multiple times in similar contexts.

2 Experiment

We constructed three successively larger corpora of 1, 2 and 4 million words by randomly selecting articles from the original "Wikicorpus" made freely available on the internet by Reese et al. (2010). Wikicorpus is itself based on articles from the collaborative encyclopedia Wikipedia. We selected the upper bound of 4 M words to be comparable with the typical corpus size used in theoretical studies on LSA (see for instance Landauer and Dumais (1997) and Griffiths et al. (2007)). For each corpus, we kept only words that occurred at least 10 times and we excluded a stop list of high frequency words with no conceptual content such as: the, of, to, and ... This left us with a vocabulary of 8 643, 14 147 and 23 130 words respectively. For the simulations, we used the free software Gensim (Řehůřek and Sojka, 2010) that provides an online Python implementation of LSA.

We first reproduced the results of Griffiths et al. (2007), from which we derived the behavior-based measure. Then, we computed our corpus-based measure with the same models.

2.1 The behavior-based measure

Following Griffiths et al. (2007), we used the free association norms collected by Nelson et al. (1998) as a gold standard to study the psychological relevance of the LSA semantic representation. The norms were constructed by asking more than 6000 participants to produce the first word that came to mind in response to a cue word. The participants were presented with 5,019 stimulus words and the responses (word associates) were ordered by the frequency with which they were named. The overlap between the words used in the norms and the vocabulary of our smallest corpus was 1093 words. We used only this restricted overlap in our experiment.

In order to evaluate the performance of LSA models in reproducing these human generated data, we used the same measure as in Griffiths et al. (2007): the median rank of the first associates of a word in the semantic space. This was done in three steps : 1) for each word cue W_c, we sorted the list of the remaining words W_i in the overlap set, based on their LSA cosine similarity with that cue: $cos(LSA(W_c), LSA(W_i))$, with highest cosine ranked first. 2) We found the ranks of the first three associates for that cue in that list. 3) We applied 1) and 2) to all words in the overlap set and we computed the median rank for each of the first three associates.

Griffiths et al. (2007) tested a set of semantic dimensions going from 100 to 700. We extended the range of dimensions by testing the following set : [2,5,10,20,30,40,50,100, 200, 300,400,500,600,700,800,1000]. We also manipulated the number of successive sentences to be taken as defining the context of a given word (document size), which we varied from 1 to 100.

In Figure 1 we show the results for the 4 M size corpus with 10 sentences long documents.

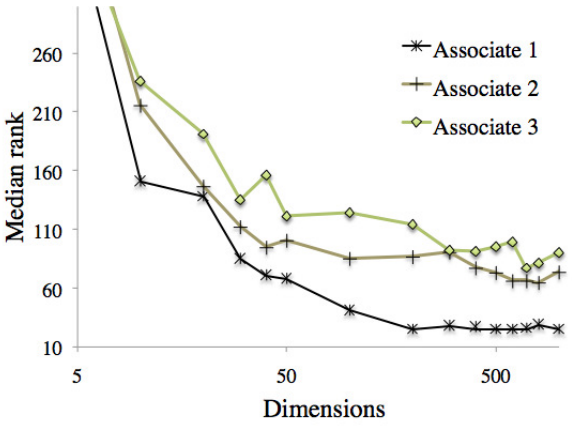

Figure 1 : The median rank of the three associates as a function of the semantic dimensions (lower is better)

For the smaller corpora we found similar results as we can see from Table 1 where the scores represent the median rank averaged over the set of dimensions ranging from 10 to 1000. As found in Griffiths et al. (2007), the median rank measure predicts the order of the first three associates in the norms.

In the rest of the article, we will need to characterize the semantic model by a single value. Instead of taking the median rank of only one of the

Size	associate 1	associate 2	associate 3
1 M	78.21	152.18	169.07
2 M	57.38	114.57	131
4 M	54.57	96.5	121.57

Table 1 : The median rank of the first three associates for different sizes

associates, we will consider a more reliable measure by averaging over the median ranks of the three associates across the overlap set. We will call this measure the Median Rank.

2.2 The Pseudo-synonym detection task

The measure we introduce in this part is based on a Same-Different Task (SDT). It is described schematically in Figure 2, and is computed as follows: for each corpus, we generate a Pseudo-Synonym-corpus (PS-corpus) where each word in the overlap set is randomly replaced by one of two lexical variants. For example, the word "Art" is replaced in the PS-corpus by "Art_1" or "Art_2". In the derived corpus, therefore, the overlap lexicon is twice as big, because each word is duplicated and each variant appears roughly with half of the frequency of the original word.

The Same-Different Task is set up as follows: a pair of words is selected at random in the derived corpus, and the task is to decide whether they are variants of one another or not, only based on their cosine distances. Using standard signal detection techniques, it is possible to use the distribution of cosine distances across the entire list of word pairs in the overlap set to compute a Receiver Operating Characteristic Curve (Fawcett, 2006), from which one derives the area under the curve. We will call this measure : SDT-ρ. It can be interpreted as the probability that given two pairs of words, of which only one is a pseudo-synonym pair, the pairs are correctly identified based on cosine distance only. A value of 0.5 represents pure chance and a value of 1 represents perfect performance.

It is worth mentioning that the idea of generating pseudo-synonyms could be seen as the opposite of the "pseudo-word" task used in evaluating word sense disambiguation models (see for instance Gale et al. (1992) and Dagan et al. (1997)). In this task, two different words w_1 and w_2 are combined to form one ambiguous pseudo-word $W_{12} = \{w_1, w_2\}$ which replaces both w_1 and w_2 in the test set.

We now have two measures evaluating the quality of a given semantic representation: The Median Rank (behavior-based) and the SDT-ρ (corpus-based). Can we use the latter to predict the former? To answer this question, we compared the performance of both measures across different semantic models, document lengths and corpus sizes.

3 Results

In Figure 3 (left), we show the results of the behavior-based Median Rank measure, obtained from the three corpora across a number of semantic dimensions. The best results are obtained with a few hundred dimensions. It is important to highlight the fact that small differences between high dimensional models do not necessarily reflect a difference in the quality of the semantic representation. In this regard, Landauer and Dumais (1997) argued that very small changes in computed cosines can in some cases alter the LSA ordering of the words and hence affect the performance score. Therefore only big differences in the Median Ranks could be explained as a real difference in the overall quality of the models. The global trend we obtained is consistent with the results in Griffiths et al. (2007) and with the findings in Landauer and Dumais (1997) where maximum performance for a different task (TOEFL synonym test) was obtained over a broad region around 300 dimensions.

Besides the effect of dimensionality, Figure 3 (left) indicates that performance gets better as we increase the corpus size.

In Figure 3 (right) we show the corresponding results for the corpus-based SDT-ρ measure. We can see that SDT-ρ shows a parallel set of results and correctly predicts both the effect of dimensionality and the effect of corpus size. Indeed, the general trend is quite similar to the one described with the Median Rank in that the best performance is obtained for a few hundred dimensions and the three curves show a better score for large corpora.

Figure 4 shows the effect of document length on the Median Rank and SDT-ρ. For both measures, we computed these scores and averaged them over the three corpora and the range of dimensions going from 100 to 1000. As we can see, SDT-ρ predicts the psychological optimal document length,

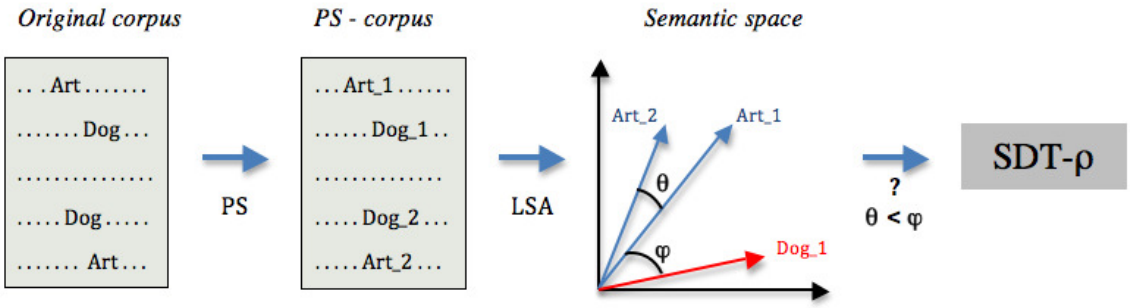

Figure 2 : Schematic description of the Same-Different Task used.

which is about 10 sentences per document. In the corpus we used, this gives on average of about 170 words/document. This value confirms the intuition of Landauer and Dumais (1997) who used a paragraph of about 150 word/document in their model.

Finally, Figure 5 (left) summarizes the entire set of results. It shows the overall correlation between SDT-ρ and the Median Rank. One point in the graph corresponds to a particular choice of semantic dimension, document length and corpus size. To measure the correlation, we use the Maximal Information Coefficient (MIC) recently introduced by Reshef et al. (2011). This measure captures a wide range of dependencies between two variables both functional and not. For functional and non-linear associations it gives a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. For our data this correlation measure yields a score of $MIC = 0.677$ with ($p < 10^{-6}$).

In order to see how the SDT-ρ measure would correlate with another human-generated benchmark, we ran an additional experiment using the TOEFL synonym test (Landauer and Dumais, 1997) as gold standard. It contains a list of 80 questions consisting of a probe word and four answers (only one of which is defined as the correct synonym). We tested the effect of semantic dimensionality on a 6 M word sized Wikipedia corpus where documents contained respectively 2, 10 and 100 sentences for each series of runs. We kept only the questions for which the probes and the 4 answers all appeared in the corpus vocabulary. This left us with a set of 43 questions. We computed the response of the model on a probe word by selecting the answer word with which it had the smallest cosine angle. The best performance (65.1% correct) was obtained with 600 dimensions. This is similar to the result reported in Landauer and Dumais (1997) where the best performance obtained was 64.4% (compared to 64.5% produced by non-native English speakers applying to US colleges). The correlation with SDT-ρ is shown in Figure 5 (right). Here again, our corpus-based measure predicts the general trend of the behavior-based measure: higher values of SDT-ρ correspond to higher percentage of correct answers. The correlation yields a score of $MIC = 0.675$ with ($p < 10^{-6}$).

In both experiments, we used the overlap set of the gold standard with the Wikicorpus to compute the SDT-ρ measure. However, as the main idea is to apply this evaluation method to corpora for which there is no available human-generated gold standards, we computed new correlations using a SDT-ρ measure computed, this time, over a set of randomly selected words. For this purpose we used the 4M corpus with 10 sentences long documents and we varied the semantic dimensions. We used the Median Rank computed with the Free association norms as a behavior-based measure.

We tested both the effect of frequency and size: we varied the set size from 100 to 1000 words which we randomly selected from three frequency ranges : higher than 400, between 40 and 400 and between 40 and 1. We chose the limit of 400 so that we can have at least 1000 words in the first range. On the other hand, we did not consider words which occur only once because the SDT-ρ requires at least two instances of a word to generate a pseudo-synonym.

The correlation scores are shown in Table 2. Based on the MIC correlation measure, mid-

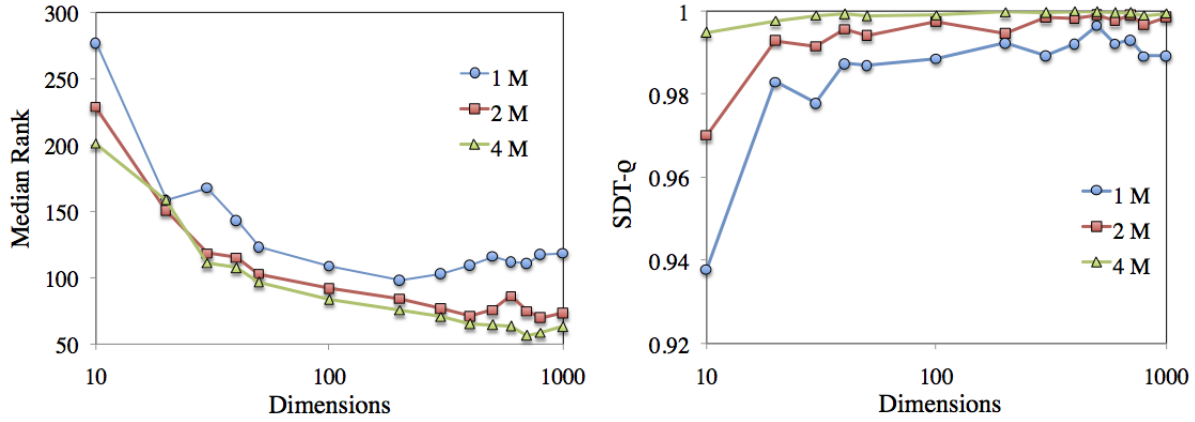

Figure 3 : The Median rank (left) and SDT-ρ (right) as a function of a number of dimensions and corpus sizes. Document size is 10 sentences.

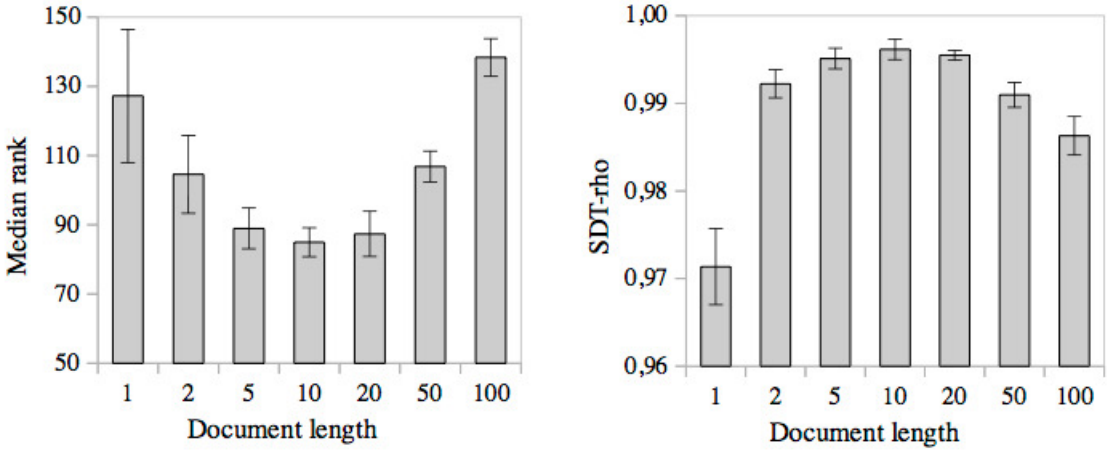

Figure 4 : The Median rank (left) and SDT-ρ (right) as a function of document length (number of sentences). Both measures are averaged over the three corpora and over the range of dimensions going from 100 to 1000.

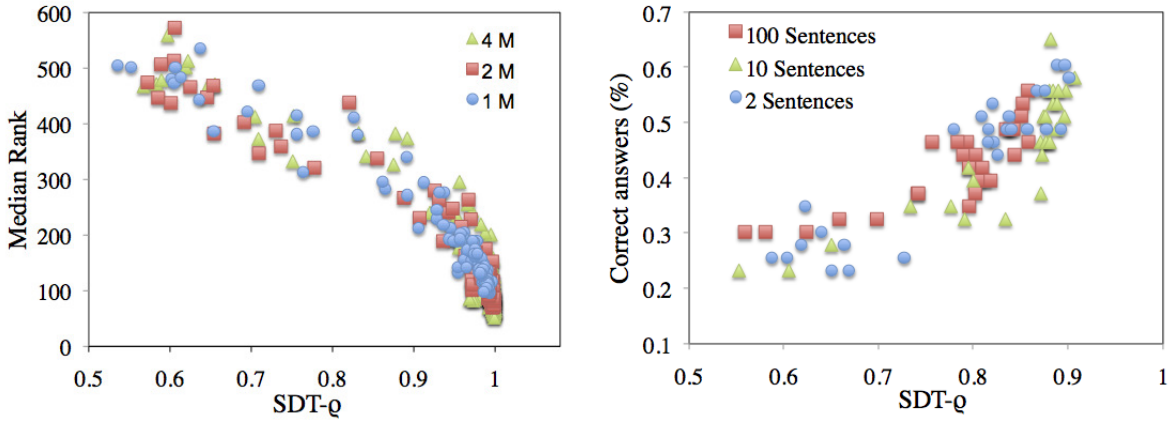

Figure 5 : Overall correlation between Median Rank and SDT-ρ (left) and between Correct answers in TOEFL synonym test and SDT-ρ (right) for all the runs. .

Freq. x	$1 < x < 40$			$40 < x < 400$			$x > 400$			**All**	**Overlap**
Size	100	500	1000	100	500	1000	100	500	1000	~ 4 M	1093
MIC	0.311	0.219	0.549^*	0.549^*	$\mathbf{0.717}^*$	$\mathbf{0.717}^*$	0.311	0.205	0.419	0.549^*	$\mathbf{0.717}^*$

$^*: p < 0.05$

Table 2 : Correlation scores of the Median Rank with the SDT-ρ measure computed over randomly selected words from the corpus, the whole lexicon and the overlap with the free association norms. We test the effect of frequency and set size.

frequency words yield better scores. The correlations are as high as the one computed with the overlap even with a half size set (500 words). The overlap is itself mostly composed of mid-frequency words, but we made sure that the random test sets have no more than 10% of their words in the overlap. Mid-frequency words are known to be the best predictors of the conceptual content of a corpus, very common and very rare terms have a weaker discriminating or "resolving" power (Luhn, 1958).

4 Discussion

We found that SDT-ρ enables to predict the outcome of behavior-based evaluation methods with reasonable accuracy across a range of parameters of a LSA model. It could therefore be used as a proxy when human-generated data are not available. When faced with a new corpus and a task involving similarity between words, one could implement this rather straightforward method in order, for instance, to set the semantic model parameters.

The method could also be used to compare the performance of different distributional semantic models, because it does not depend on a particular format for semantic representation. All that is required is the existence of a semantic similarity measure between pairs of words. However, further work is needed to evaluate the robustness of this measure in models other than LSA.

It is important to keep in mind that the correlation of our measure with the behavior-based methods only indicates that SDT-ρ can be trusted, to some extent, in evaluating these semantic tasks. It does not necessarily validate its ability to assess the entire semantic structure of a distributional model. Indeed, the behavior-based methods are dependent on particular tasks (i.g., generating associates, or responding to a multiple choice synonym test) hence they represent only an indirect evaluation of a model, viewed through these specific tasks.

It is worth mentioning that Baroni and Lenci (2011) introduced a comprehensive technique that tries to assess simultaneously a variety of semantic relations like meronymy, hypernymy and coordination. Our measure does not enable us to assess these relations, but it could provide a valuable tool to explore other fine-grained features of the semantic structure. Indeed, while we introduced SDT-ρ as a global measure over a set of test words, it can also be computed word by word. Indeed, we can compute how well a given semantic model can detect that "Art_1" and "Art_2" are the same word, by comparing their semantic distance to that of random pairs of words. Such a word-specific measure could assess the semantic stability of different parts of the lexicon such as concrete vs. abstract word categories, or the distribution properties of different linguistic categories (verb, adjectives, ..). Future work is needed to assess the extent to which the SDT-ρ measure and its word-level variant provide a general framework for DSMs evaluation without external resources.

Finally, one concern that could be raised by our method is the fact that splitting words may affect the semantic structure of the model we want to assess because it may alter the lexical distribution in the corpus, resulting in unnaturally sparse statistics. There is in fact evidence that corpus attributes can have a big effect on the extracted model (Sridharan and Murphy, 2012; Lindsey et al., 2007). However, as shown by the high correlation scores, the introduced pseudo-synonyms do not seem to have a dramatic effect on the model, at least as far as the derived SDT-ρ measure and its predictive power is concerned. Moreover, we showed that in order to apply the method, we do not need to use the whole lexicon, on the contrary, a small test set of about 500 random mid-frequency words (which represents less than 2.5 % of the total vocabulary) was shown to lead to better results. However, even if the results are not directly affected in our case, future work needs to investigate the exact effect word splitting may have on the semantic model.

References

Andrews, M., G. Vigliocco, and D. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review 116*, 463–498.

Baroni, M. and A. Lenci (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop*, East Stroudsburg PA: ACL, pp. 1–10.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM 55*(4), 77–84.

Dagan, I., L. Lee, and F. Pereira (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th ACL/8th EACL*, pp. 56–63.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Gale, W., K. Church, and D. Yarowsky (1992). Work on statistical methods for word sense disambiguation. *Workings notes, AAAI Fall Symposium Series, Probabilistic Approaches to Natural Language*, 54–60.

Griffiths, T., M. Steyvers, and J. Tenenbaum (2007). Topics in semantic representation. *Psychological Review 114*, 114–244.

Harris, Z. (1954). Distributional structure. *Word 10*(23), 146–162.

Hutchison, K., D. Balota, M. Cortese, and J. Watson (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology 61*(7), 1036–1066.

Kievit-Kylar, B. and M. N. Jones (2012). Visualizing multiple word similarity measures. *Behavior Research Methods 44*(3), 656–674.

Landauer, T. and S. Dumais (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Lewandowsky, S. and S. Farrell (2011). *Computational modeling in cognition : principles and practice*. Thousand Oaks, Calif. : Sage Publications.

Lindsey, R., V. Veksler, and A. G. andWayne Gray (2007). Be wary of what your computer reads: The effects of corpus selection on measuring semantic relatedness. In *Proceedings of the Eighth International Conference on Cognitive Modeling*, pp. 279–284.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development 2*(2), 157–165.

Nelson, D., C. McEvoy, and T. Schreiber (1998). The university of south florida word association, rhyme, and word fragment norms.

Reese, S., G. Boleda, M. Cuadros, L. Padro, and G. Rigau (2010). Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)*, La Valleta, Malta.

Řehůřek, R. and P. Sojka (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50.

Reshef, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti (2011). Detecting novel associations in large datasets. *Science 334*(6062), 1518–1524.

Sridharan, S. and B. Murphy (2012). Modeling word meaning: distributional semantics and the sorpus quality-quantity trade-off. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, COLING 2012, Mumbai, pp. 53–68.

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