Improving a Fundamental Measure of Lexical Association

Gabriel Recchia (glr29@cam.ac.uk) **Paul Nulty (pgn26@cam.ac.uk)**

Cambridge Centre for Digital Knowledge, CRASSH, Alison Richard Building, 7 West Rd
University of Cambridge, Cambridge CB3 9DP, United Kingdom

Abstract

Pointwise mutual information (PMI), a simple measure of lexical association, is part of several algorithms used as models of lexical semantic memory. Typically, it is used as a component of more complex distributional models rather than in isolation. We show that when two simple techniques are applied—(1) down-weighting co-occurrences involving low-frequency words in order to address PMI's so-called "frequency bias," and (2) defining co-occurrences as counts of "*events in which instances of word$_1$ and word$_2$ co-occur in a context*" rather than "*contexts in which word$_1$ and word$_2$ co-occur*"—then PMI outperforms default parameterizations of word embedding models in terms of how closely it matches human relatedness judgments. We also identify which down-weighting techniques are most helpful. The results suggest that simple measures may be capable of modeling certain phenomena in semantic memory, and that complex models which incorporate PMI might be improved with these modifications.

Keywords: semantic spaces; word space models; semantic memory; semantic networks; computational models

Introduction

Pointwise mutual information (PMI) is a simple measure that plays an important role in many computational models that approximate human judgments of lexical association or semantic relatedness. Such "semantic space" models typically take the form of algorithms that process a corpus of written language, such as Wikipedia or TASA, and construct quantitative representations of the words they encounter on the basis of lexical co-occurrence statistics. The resulting 'lexical representations' (e.g., numerical vectors) are intended to correspond roughly to semantic representations in the human mind, at least at some level of abstraction. Of particular interest is the *degree of association* that exists between related (and unrelated) words in any such model. This quantity is computed in a manner appropriate to the model at hand, e.g. cosine similarity between two lexical vectors in a vector space model, or Kullback–Leibler divergence between distributions of words over topics in a topic model. Such computationally estimated associations can then be compared to behavioral data that provides evidence of the actual degree to which people perceive particular words to be related, e.g., human judgments of the semantic relatedness of large numbers of word pairs.

Such correlations with behavioral data are frequently used to argue in favor of particular models of human semantic memory (Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Bullinaria & Levy, 2007), but lexical associations derived from semantic space models have many other applications as well. For example, a range of semantic space models—including one method that has been recently shown by Levy & Goldberg (2014) to be implicitly factorizing a matrix of PMI scores—have recently been employed to study associative processing in high-level judgment, modeling phenomena such as the conjunction fallacy and naturalistic judgment problems (Bhatia, 2017). PMI or explicitly PMI-based methods have been used to cluster terms syntactically and semantically (Bullinaria & Levy, 2007, 2012), recognize synonyms (Turney, 2001), automatically identify clusters that correspond to different senses of a word's meaning (Pantel & Lin, 2002), extract linguistic collocations from text (Manning & Schütze, 1999), and identify patterns of relationships between symptoms in dementia (Mitnitski, Richard, & Rockwood, 2014), among many other applications.

Because of the range of applications to which PMI and PMI-based methods are applied, any modifications that improved PMI's ability to model human semantic judgments would potentially have benefits for the wide range of computational methods in which it is a component. Furthermore, if a slight modification of some neurally plausible algorithm such as PMI was to produce lexical associations that were as good as those produced by state-of-the-art models (in terms of correlation to human data), it would be worth investigating as a possible computational simplification/abstraction of some process actually taking place within human semantic memory. Finally, simple, computationally efficient yet accurate means of estimating lexical associations are useful within the field of artificial intelligence, as they can more readily be scaled up to larger datasets than can methods that take longer to compute. For all of these reasons, simple measures of lexical association are worthy of closer investigation.

PMI is traditionally defined as follows (Church & Hanks, 1989):

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

This formulation "compares the probability of observing *x* and *y together* (the joint probability) with the probabilities of observing *x* and *y independently* (chance)" (Church & Hanks, 1989, p. 77). Estimating these probabilities is commonly done in a straightforward manner: P(*x*,*y*) is estimated by dividing the number of "contexts" (documents, windows of text, etc.) in which *x* and *y* co-occur by the total number of contexts in the corpus, and P(*x*) is estimated by

dividing the number of contexts containing *x* by the total number of contexts in the corpus (and likewise for P(*y*)) (Manning & Shütze, 1999; Turney & Pantel, 2010).

Strengths and Weaknesses of PMI

PMI is a component of many different algorithms that have been fit to behavioral data in the psychological literature. For example, a slight variant of it (PPMI, or 'positive PMI,' which differs only in that negative values are set to zero) has been used directly in lexical vector components in models such as 'PPMI Cosines' (Bullinaria & Levy, 2007, 2012), and as a preprocessing step to be applied to a matrix prior to singular value decomposition or other matrix factorization techniques. Some algorithms that initially seemed to have little to do with PMI are more linked to it than they first appeared. For example, consider the SGNS algorithm of the popular word embedding tool *word2vec* (Mikolov, Sustskever, Chen, Corrado, & Dean, 2013), which has been recently used in studies of metaphor perception and associative processing (Agres et al., 2016; Bhatia, 2017), and is responsible for the Google Word2Vec dataset recently described as one of several "data sets with potential relevance for cognitive science" in a recent survey (Goldstone & Lupyan, 2015, Table 2). Although this algorithm is typically conceived of as a shallow neural network, its core mathematical operations have been shown to be implicitly factorizing the "well-known word-context PMI matrix from the word-similarity literature, shifted by a constant offset" (Levy & Goldberg, 2014, p. 2177). The same appears to be true of an alternative embedding method known as noise-contrastive estimation (Levy & Goldberg, 2014). In fact, much of the advantage that "prediction-based[1]" models such as *word2vec*'s SGNS initially seemed to hold over more traditional distributional models (Baroni, Dinu, & Kruszewski, 2014) appears to be due to *word2vec*'s exploitation of 'hyperparameters' –i.e., miscellaneous operations such as smoothing and subsampling (Levy, Goldberg, & Dagan, 2015). When these more traditional vector space models are enhanced with analogous hyperparameters, they tend to do as well as prediction-based models (Levy et al., 2015).

Given the ubiquity of PMI in computational models of semantic relatedness, it seems that this measure must be capturing something important. Yet the measure is well-known for its weaknesses. The most fundamental of these is "frequency bias," PMI's tendency to over-weight co-occurrences involving low-frequency words (Levy et al., 2015; Manning & Shütze, 1999; Turney & Pantel, 2010). One way to think about the cause of this problem is that although probability estimates are more accurate when they are made on the basis of lots of data (e.g., frequent words) than on sparse data (infrequent words), the formula for PMI does not account for this fact. On the contrary, the less frequent the words, the lower the denominator and the larger the result. Thus a chance co-occurrence between two rare words that each occur only once in a large corpus will result in an exceedingly high PMI. Because Zipf's law entails that any corpus will have many more infrequent than frequent lexical types, this problem is pervasive.

Addressing PMI's Weaknesses

Given the fundamental difficulties inherent in estimating co-occurrence probabilities from infrequent words, various adjustments to PMI have been proposed to mitigate the problem. Here we consider one commonly proposed solution (down-weighting co-occurrences involving low-frequency words in some way, to counter PMI's tendency to over-weight them), and one solution that we have not previously seen proposed (adjusting how 'co-occurrences' are defined/counted).

Down-weighting. The probabilities in the denominator of the PMI formula naturally down-weight co-occurrences involving frequent words. This is a desired property; without the denominator, the most "associated" words with virtually any term would be "the," "of," and many other words that occur very frequently across the board. As previously mentioned, however, PMI (and PPMI) have the opposite problem, in that the words these measures deem to be most semantically related to a word *w* "are often extremely rare words, which do not necessarily appear in the respective representations of words that are semantically similar to *w*" (Levy et al., 2015, p. 213).

As such, several modifications to PMI have been proposed, many of which are enumerated in Table 1. The ultimate goal of all of these is to cause low frequency words to be ranked less highly than in the standard PMI formula. Some further adjustments have been proposed which rely on information other than the co-occurrence counts and frequencies of the words whose association is being calculated. Because these rely on additional information, there is sometimes a fine line between such modifications of PMI and novel distributional models, and they often have additional parameters. Yet other measures, such as PMI^2 , have been shown to be monotonic transformations of other measures already appearing in Table 1 (Evert, 2005). We confine our comparisons in Study 1 to only the simplest measures, i.e., measures that, when computing the degree of association between words w_1 and w_2, rely only upon the corpus-wide counts $f(w_1)$ and $f(w_2)$, and the co-occurrence counts $f(w_1, w_2)$.

Counting. It is clear that there is variation in the literature with respect to the manner in which the probabilities involved in PMI are estimated. For example, several researchers report estimating P(*x*) the as frequency of *x* divided by the number of words in the corpus (Church & Hanks, 1989; Islam & Inkpen, 2008), while others use the number of documents in which *x* appears divided by the

[1] These are distributional semantic models that "frame the vector estimation problem directly as a supervised task, where the weights in a word vector are set to maximize the probability of the contexts in which the word is observed in the corpus" (Baroni et al., 2014, p. 238.)

number of documents in the corpus (Manning & Shütze, 1999; Turney, 2001; Turney & Pantel, 2010). Similarly, many authors mention that they use the "number of co-occurrences" of *x* and *y* to estimate P(*x*, *y*), without specifying exactly what counts as a "co-occurrence." A reasonable assumption is that in some cases this is shorthand for "the number of contexts/documents in which *x* and *y* co-occur," and indeed this seems to be approach of some authors who spell out their calculations in detail (Manning & Shütze, 1999; Turney, 2001; Turney & Pantel, 2010). A more literal interpretation of "number of co-occurrences"—and perhaps the one intended by at least some of the authors who have used this phrase—would be that this refers to the number of co-occurrence *events*. For example, in the sentence context "Tiger, tiger, burning bright," the word type *tiger* can be conceived of as co-occurring with *bright* twice (one co-occurrence for each instance of *tiger*)[2,3]. We will refer to this method of co-occurrence counting as "event-based counting," as contrasted from the "context-based" method of counting the number of contexts/documents in which *x* and *y* appear together. Event-based counting attends to the information available in the corpus at a more fine-grained level than does context-based counting, as it distinguishes between contexts in which word pairs might appear many times and contexts in which they might appear together only a single time. As such, it can be seen as increasing the overall amount of evidence about word associations that go into the estimation of the probabilities.

Study 1

Down-weighting and event-based counting each have the potential to address PMI's frequency bias—the former by compensating for the fact that rarer words provide weaker evidence, and the latter by bolstering the overall amount of evidence that the measure takes into account. In Study 1, the success of each approach is evaluated individually and in combination. Table 1 provides the formulae for each of the down-weighting methods surveyed in the previous section, with citations provided in footnotes. Some methods, namely SCI, SCI_{sig}, and context distribution smoothing, are asymmetric and distinguish between a cue word *x* and a response word *y*.

In theory, either context-based or event-based counting could be used with any one of these measures. With context-based counting, *P(x, y)* is estimated by dividing the total number of contexts in which *x* and *y* appear together by a constant factor, namely the total number of contexts in the corpus (Turney & Pantel, 2010). Analogously, with event-based counting, it makes sense to divide the number of co-occurrence events in which *x* and *y* appear together by the total number of co-occurrence events in the corpus $\sum_i^N |context_i|(|context_i| - 1)$. In practice, however, the specific value here is irrelevant, as it merely serves to scale all PMI scores by a constant factor.

Analogously, to estimate the 'global' or 'corpus-wide' probability P(*x*) of observing a word, we can either count the total number of contexts in which *x* appears (context-based counting), or we can count *x*'s raw frequency – the total number of times x appears anywhere in the corpus (event-based counting), and divide the result by the relevant constant factor (number of contexts, or number of co-occurrence events).

Some of the measures in Table 1 call for the use of co-occurrence frequencies *f(x,y)* or global frequencies (*f(x)*, *f(y)*). These are counted as previously described, except that they are not divided by a constant factor.

Table 1: Methods for down-weighting PMI scores.

Method	Formula
"Discount factor"[4]	$\left(\frac{f(x,y)}{f(x,y)+1}\right)\left(\frac{\min(f(x),f(y))}{\min(f(x),f(y))+1}\right)pmi$
SCI[5]	$\frac{P(x,y)}{P(x)\sqrt{P(y)}}$
PMI_{sig}[5]	$\sqrt{\min(P(x),P(y))}\left(\frac{P(x,y)}{P(x)P(y)}\right)$
SCI_{sig}[5]	$\sqrt{\min(P(x)P(y))}\left(\frac{P(x,y)}{P(x)\sqrt{P(y)}}\right)$
gmean[6]	$\frac{f(x,y)}{\sqrt{f(x)f(y)}}$
Context distribution smoothing[7]	$log\left(\frac{P(x,y)}{P(x)\frac{f(y)^\alpha}{\sum_i f(i)^\alpha}}\right)\ with\ \alpha = 0.75$

Method

Word pair lists were obtained for all semantic relatedness tasks evaluated in Recchia and Jones (2009), namely the tasks of Miller & Charles (1991), Resnik (1995), Rubenstein & Goodenough (1965), and Finkelstein et al. (2002). Because the latter task conflates judgments of semantic similarity (*{car, truck}*) with judgments of semantic relatedness (*{car, road}*), we used the version of this task that had been partitioned into the so-called "WordSim Similarity" and "WordSim Relatedness" subsets (Agirre et al., 2009). Also included was an additional similarity task, SimLex-999 (Hill, Reichart, & Korhonen, 2014) and two additional relatedness tasks referred to in the literature as

[2] Co-occurrences are generally viewed as symmetric relations, and we will keep with that tradition here: *tiger* co-occurs with *bright* twice in this sentence, and vice versa.

[3] This is the approach of Church & Hanks (1989) and Islam & Inkpen (2008), except that their contexts are defined as windows of text (i.e., strings containing *n* words); the size of the window is an additional parameter for the model.

[4] Pantel & Lin (2002)

[5] Washtell & Markert (2009)

[6] Evert (2005)

[7] Levy, Goldberg, & Dagan (2015)

MEN (Bruni, Boleda, Baroni, & Tran, 2012) and MTurk (Radinsky, Agichtein, Gabrilovitch, & Markovitch, 2011).

Raw PMI scores as well as each of the down-weighting metrics in Table 1 were calculated for every word pair in each relatedness and similarity task[8], using a version of the Westbury Lab Wikipedia Corpus (Shaoul & Westbury, 2010) with punctuation removed and capital letters converted to lower case. The resulting corpus contained 3,035,070 documents and approximately 1 billion words. Each metric was computed with context-based counting as well as with event-based counting as described in detail on the previous page. Rather than a window size, terms were treated as 'co-occurring' if they appeared in the same document (i.e., Wikipedia article).

Additionally, to get a sense of how these metrics stack up against what are perhaps the most popular distributional models today—the *word2vec* CBOW and SGNS models—we trained each *word2vec* model on the same corpus using the default settings recommended by Google[9], and used the resulting vectors to estimate semantic relatedness in the standard manner (e.g., computing cosines between 300-dimensional vectors). Comparing to distributional models whose parameters have not been optimized for the tasks at hand is in some ways an unfair comparison. Nevertheless, *word2vec*'s 'off-the-shelf' parameters are the ones most frequently employed when *word2vec* is used in real-world settings. As usual, Spearman rank correlations were computed between each metric and the human judgments provided by each relatedness and similarity task.

Results

Down-weighting methods. The only down-weighting methods tested that were consistently as good as or better than the standard PMI formula were the discount factor of Pantel & Lin (worse performance than raw PMI on 1 of the 8 tasks when using context-based counts, 2 tasks when when using event-based counts) and the "context distribution smoothing" of Levy et al. (worse performance than raw PMI on only 1 task, irrespective of counting method employed). All other down-weighting methods exhibited worse performance than raw PMI on over half of all tasks regardless of counting method. Table 2 illustrates Spearman rank correlations between human judgments and these best-performing down-weighting methods using context-based counting, event-based counting, and the two *word2vec* models.

Counting methods. Restricting ourselves to the down-weighting methods that produced reliable improvements, event-based counting resulted in higher correlations to human data than did context-based counting on all tasks except for SimLex-999. Across all tasks, using event-based rather than context-based counting increased correlations by an average of 2.7 points for context distribution smoothing, 4.2 points for the discount factor, and 4.9 points for raw PMI scores.

Table 2: Correlations with human judgments of semantic relatedness (tasks 1-5, 7) and similarity (6, 8).

	Task number (see *Note* below)							
Method	1	2	3	4	5	6	7	8
CDS, Context	.68	.75	.58	.83	.82	.32	.64	.73
DF, Context	.63	.75	.51	.85	.81	.30	.57	.66
PMI, Context	.62	.74	.50	.84	.78	.30	.57	.66
CDS, Event	**.72**	**.81**	.58	**.87**	**.86**	.27	**.68**	**.76**
DF, Event	.70	.79	.55	.86	.83	.29	.66	.72
PMI, Event	.70	.79	.55	.86	.82	.29	.66	.72
SGNS	.71	.77	**.64**	.82	.75	.30	.62	.75
CBOW	.67	.71	.56	.73	.67	**.32**	.47	.72

Note. CDS: context distribution smoothing, DF: discount factor; PMI: unmodified PMI; SGNS: *word2vec* skip-grams with negative sampling; CBOW: *word2vec* 'continuous bag of words'; "Context" and "Event" refer to the counting method used. Task numbers refer to the judgments of semantic relatedness/similarity compiled by 1: Bruni et al. (2012); 2: Miller & Charles (1991); 3: Radinsky et al. (2011); 4: Resnik (1995); 5: Rubenstein & Goodenough (1965); 6: Hill et al. (2014); 7: WordSim-Relatedness (Agirre et al., 2009); 8: WordSim-Similarity (Agirre et al., 2009). The highest correlation for each task appears in bold.

Discussion

Down-weighting and event-based smoothing both confer advantages when PMI is used to estimate semantic relatedness judgments. Specifically, the combination of context distribution smoothing (CDS) and event-based counting performed best for all datasets except for two. Each of these was a dataset on which the various versions of PMI all performed poorly. When SimLex-999 was constructed (Hill et al., 2014), respondents were given explicit instructions about the difference between similarity and relatedness, and told to judge similarity only. PMI has no mechanism for distinguishing between related and similar terms, and does not detect relationships between paradigmatically related terms (which tend to be *similar*) as well as SGNS does. It is not clear why all metrics did well on WordSim-Similarity, but one reason may be that Agirre et al. (2009) did not specifically instruct participants to rate word pairs based on their similarity. Rather, they created WordSim-Similarity with the original judgments from Finkelstein et al. (2002), which had instructions that conflated relatedness and similarity, but they excluded related word pairs that did not share a formal similarity relation (synonymy, antonymy, hyponymy, etc.)

Why does context distribution smoothing work? Given that the $\sum_i f(i)^\alpha$ term is constant for any fixed value of α,

[8] If computing a metric resulted in an undefined value (*log 0*), the value of the metric was replaced with zero.

[9] That is, a window size of 10 for SGNS and 5 for CBOW (as recommended at https://code.google.com/archive/p/word2vec/), and all other parameters left on their default settings.

the only thing that really seems to distinguish CDS from the other discounting methods is its use of α (set to .75) in the exponent of *f(y)*. Furthermore, since P(*y*) is estimated by dividing *f(y)* by another constant, context distribution smoothing is closely related to the much more poorly performing SCI metric of Washtell & Markert (2009), $\frac{P(x,y)}{P(x)\sqrt{P(y)}}$, which merely raises P(*y*) to the power of .5 rather than .75.

Why would there be anything special about .75? One possibility is that this value strikes the proper balance between raising P(*y*) to the value of 0 (which would ignore the frequency of P(*y*) and result in a measure that was highly correlated with *y*'s frequency), versus raising P(*y*) to the value of 1 (yielding PMI, which is known to give outsize values to infrequent words and is thus likely inversely correlated with frequency). In other words, down-weighting may be optimal when it yields a measure that is neither positively nor negatively correlated with word frequency. This possibility is briefly explored in Study 2.

Study 2

To find the α for which CDS yields a correlation with word frequency as close to zero as possible, α was fit so as to minimize the absolute value of the Spearman rank correlation between word frequency[10] and CDS. Because there is no reason in this context to modify P(y) but not P(x), the same was done for a generalization of the *gmean* measure, "simple" smoothing, defined simply as $log\left(\frac{P(x,y)}{P(x)^{\alpha}P(y)^{\alpha}}\right)$. Finally, the value of α that maximized correlations to human data was determined for both measures. Event-based counting was used in all cases due to its superiority over context-based counting in Study 1.

Table 3 illustrates the values of α that minimized the absolute value of the correlation between the measure and word frequency, while Table 4 shows values of α that maximized correlations with human judgments.

Table 3: Values of α that minimized absolute value of correlations with word frequency (Study 2)

	Task number (see *Note* below Table 2)							
Measure	1	2	3	4	5	6	7	8
CDS	.85	.77	1.0	.76	.78	.74	.77	.72
Simple	.91	.82	1.0	.79	.85	.84	.84	.82

Table 4: Values of α that maximized correlations with human judgments (Study 2)

	Task number (see *Note* below Table 2)							
Measure	1	2	3	4	5	6	7	8
CDS	.77	.80	.52	.80	.74	.97	.76	.74
Simple	.85	.81	.76	.81	.81	1.0	.84	.87

[10] Specifically, the frequency of the lowest-frequency word in each word pair.

Discussion

For CDS, the values of alpha that minimized the absolute value of the measure's correlation to word frequency (median .77) were not far off from the values of alpha that maximized correlations to human judgments (median .765), with the greatest discrepancies being on those tasks on which CDS did not perform well in Study 1 (#3 and #6). The same was true of simple smoothing (medians .84 and .825, respectively). This suggests that explicitly finding ways to minimize the degree to which lexical measures of association are confounded with word frequency and other covariates could be a promising path toward improving their ability to model human data. Other future directions could include more in-depth exploration of why α so closely corresponds to those values that maximized correlations to human judgements. For example, if an experimental study showed that the same was true of study participants making judgments about the relatedness of words in an artificial language, even when this value was not equal to .75, this would provide better evidence that the human mind employs some process that makes an explicit correction for low-frequency events analogous to that proposed by CDS.

It should not be concluded from the results of Studies 1 and 2 that PMI is more effective in isolation than distributional models such as *word2vec*. It should also be noted that not all datasets are independent. For example, the word pairs in Miller & Charles (1991) and Resnik (1995) are subsets of Rubenstein & Goodenough (1965), so it is unsurprising that a measure that does well on one would do well on all three. Even so, the fact that the use of event-based counting and CDS down-weighting causes PMI to generally outperform *word2vec* on its default settings suggests that PMI may be a better abstraction of human relatedness judgments than it is commonly understood to be. Furthermore, given that PMI has so many different applications within cognitive science and is a component of so many models of lexical processing, any improvements to this measure have the potential to improve model fits across a wide range of computational studies of cognition.

Acknowledgments

The authors gratefully acknowledge the support of the Concept Lab and the Cambridge Centre for Digital Knowledge (CCDK) at the University of Cambridge.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M. & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pgs. 19–27). Stroudsburg, PA: ACL.

Agres, K. R., McGregor, S., Rataj, K., Purver, M., & Wiggins, G. A. (2016). Modeling metaphor perception with distributional semantics vector space models. In

Workshop on Computational Creativity, Concept Invention, and General Intelligence. *Proceedings of 5th International Workshop, C3GI at ESSLI* (pp. 1-14). New York: Springer.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the Association for Computational Linguistics* (pp. 238-247). Stroudsburg, PA: ACL.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review, 124*(1), 1-20.

Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in technicolor. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 1 (pp. 136–145). Stroudsburg, PA: ACL.

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.

Church, K. W. & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Proceedings of the Association for Computational Linguistics* (pp. 76-83). Stroudsburg, PA: ACL.

Evert, S. (2005). The statistics of word cooccurrences: Word pairs and collocations. PhD thesis, IMS Stuttgart.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. ACM Transactions on Information Systems, 20, 116-131.

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science, 8*(3), 548-568.

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

Islam, A. & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data, 2*(2), 10:1-25.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1-37.

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

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics, 3*, 211-225.

Manning, C. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

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

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. Language & Cognitive Processes, 6, 1-28.

Mitnitski, A., Richard, M., & Rockwood, K. (2014). Network visualization to discern patterns of relationships between symptoms in dementia. *Alzheimer's & Dementia, 10*(4), P752-P753.

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 613-619). New York: ACM.

Radinsky, K., Agichtein, E., Gabrilovich, E. & Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on the WWW* (pgs. 337–346). New York: ACM.

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.

Resnik, P. (1995). Using information content to evaluate semantic similarity. In C. S. Mellish (Ed.), Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 448-453). San Francisco: Morgan Kaufmann.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. Communications of the ACM, 8, 627-633.

Shaoul, C. & Westbury C. (2010). The Westbury Lab Wikipedia Corpus. Edmonton, AB: University of Alberta. http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning (pp. 491-502). Berlin: Springer.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*(1), 141-188.

Washtell, J., & Markert, K. (2009). A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 628-637). Stroudsburg, PA: ACL.