The Bidirectional Co-occurrence Measure: A Look at Corpus-based Co-occurence Statistics and Current Test Sets

Nicolas Neubauer, Nils Haldenwang, Oliver Vornberger

University of Osnabrück, Institute of Computer Science
Albrechtstraße 28, 49076 Osnabrück, Germany
{nneubauer, nils.haldenwang, oliver}@uni-osnabrueck.de

Abstract

Determining the meaning of or understanding a given input in natural language is a challenging task for a computer system. This paper discusses a well known technique to determine semantic relatedness of terms: corpus-based co-occurrence statistics. Aside from presenting a new approach for this technique, the Bidirectional Co-occurrence Measure, we also compare it to two well-established measures, the Pointwise Mutual Information and the Normalized Google Distance. Taking these as a basis, we discuss a multitude of popular test sets, their test methodology, strengths and weaknesses while also providing experimental evaluation results.

1. Introduction

Basic text analysis algorithms often operate with so called *bag of word* models. A given document is represented by a vector containing the number of occurrences for each word. Since this creates a vector-space representation of a given document, techniques such as defining similarity measures based on mathematical distance measures in vector-space can be used for information retrieval. However, these representations are somewhat sparse, not taking into account the *meaning* of each word.

In a such a model, the words "rain" and "thunder" occurring in two different documents would not be related in any way, while they actually both belong to certain semantic concepts such as "weather". Thus, without semantic knowledge, these two documents would be, according to, for example, their *cosine similarity*, less related or even not related at all. With knowledge about the semantic relatedness of terms this could be overcome, for example by augmenting the originally sparse vectors with related concepts or using a vector similarity measure that takes into account the semantic distance of terms. An example for an information retrieval scenario that benefited greatly from semantic knowledge can for example be found in Finkelstein et al. (2001).

Multiple so called *Measures of Semantic Relatedness* (MSRs) have been introduced and evaluated in the past (Landauer and Dumais, 1997; Turney, 2001; Cilibrasi and Vitanyi, 2007; Lindsey et al., 2007; Landauer and Dumais, 1997). A common problem is the comparability of the results. The authors are using many different corpora to acquire the underlying knowledge base and evaluate their algorithms with different benchmarks (Nelson et al., 2004; Landauer and Dumais, 1997; Miller and Charles, 1991; Rubenstein and Goodenough, 1965; Baroni and Lenci, 2011), which are each designed to test specific parts of MSRs and are suffering from some drawbacks (Baroni and Lenci, 2011).

In this paper we summarize the general idea behind MSRs as well as the methodology of calculating their values based on the used corpus and other parameters. We then discuss two often used MSRs, the *Pointwise Mutual Information* (PMI) (Turney, 2001) and *Normalized Google Distance* (NGD) (Cilibrasi and Vitanyi, 2007) and compare those to a new and intuitive MSR we introduce in this paper called *Bidirectional Co-occurrence Measure* (BCM).

2. Measures of Semantic Relatedness

When talking about measuring "semantic relatedness", we should first define what we mean by saying that two terms are *semantically related*. While terms can be related in numerous ways, for example linguistically, such as synonyms or meronyms, or rather informally, such as "rain" and "thunder", we would like to look at their relations with regard to the *meaning* of the two terms.

Although the concept of how two things are related to each other is easy to understand, a very formal definition is hard to give. As El-Yaniv and Yanay (2012) quote from Wikipedia, semantic relatedness is about "How much does term A have to do with term B".

Before we are going to present three measures, that try to achieve measuring "semantic relatedness", we would like to point out that in this paper we are only looking at measures on the basis of corpus-based co-occurrence statistics. The simple idea behind this, suggested by Landauer and Dumais (1997), is that "words with similar meanings will tend to occur in similar contexts" (Bullinaria and Levy, 2007). Having provided some examples for related work in the introduction, we would like to point to said publication by Bullinaria & Levy for a more in-depth review of previous work in this field. For our further discussions we specifically chose PMI as well as NGD because of their general popularity and similarity to BCM in representing relatedness values.

2.1. Pointwise Mutual Information

The PMI (Turney, 2001) has often been used as MSR (Lindsey et al., 2007; Turney et al., 2010). The measure is based on the probability of two terms occurring in the same window of text versus the probabilities of the two terms occurring separately. The score of relatedness of two given terms x and y is computed as

$$PMI(x,y) = \log_2 \left(\frac{p(x,y)}{p(x) \cdot p(y)} \right)$$

where $p(x), p(y)$ is the probability of the the term x, respectively y, to occur in a text-window and $p(x, y)$ denotes the dependent probability that both terms x and y occur within the same text-window. When x and y are statistically independent, the probability of their co-occurrence is $p(x) \cdot p(y)$. If they are not independent, which means they often occur together, $p(x, y)$ will be greater than $p(x) \cdot p(y)$. Hence the ratio $\frac{p(x,y)}{p(x) \cdot p(y)}$ will yield higher values for dependent terms x and y and therefore PMI is a measure for the statistical dependence of the term's co-occurrence.

2.2. Normalized Google Distance

NGD (Cilibrasi and Vitanyi, 2007) is another MSR, performing well for various tasks. Like PMI it harnesses the probability of co-occurrences. NGD is computed as follows

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where $f(x), f(y)$ denotes the number of text-windows including the term x, respectively y, $f(x, y)$ the number of windows including both, and M is the overall number of text-windows.

By this definition the NGD measures the distance between two given terms. As we want to measure the relatedness, we are using the *Normalized Similarity Score* (NSS) (Lindsey et al., 2007), which is computed as

$$NSS(x, y) = 1 - NGD(x, y).$$

2.3. Bidirectional Co-occurrence Measure

The Bidirectional Co-occurrence Measure is based on an intuitive usage of the raw co-occurrence frequencies combined with standard weighting techniques and normalization:

Following the idea that a higher number of co-occurrences within the same context correlates with a stronger relatedness of two words, we are still facing the problem that single words frequencies follow Zipf's Law and thus for example $f(a,$"the"$)$ will outrank almost every other combination of words. To overcome this, we chose to leverage the *Inverse Document Frequency (idf)* of words (Jones, 1972), which is a common weighting technique in information retrieval, in combination with the raw co-occurrence counts.

We then define the *directed strength* from term a to b, with $f(a, b)$ being the raw co-occurrence frequency, as $d(a, b) = f(a, b) \cdot idf(b)$ and the *normalized directed strength* with T being the set of all terms in the corpus as

$$dn(a, b) = \frac{d(a, b)}{\max_{t \in T} d(a, t)}.$$

Now, $dn(a, b)$ is an intuitive measure of how important term b is in the context of term a with values in $[0, 1]$. The final measure takes the "importance" of b in the context of a and multiplies it with the "importance" of a in the context of b: $bcm(a, b) = dn(a, b) \cdot dn(b, a)$. This not only amplifies ratings for words which are important in each others context but also further damps the final ranking of frequent words since $dn(a, b)$ is normalized: Although $dn(a,$"the"$)$ might still be quite high, $dn($"the"$, a)$ probably is not.

3. Methodology

Research indicates that corpus selection significantly influences the performance of MSRs. Lindsey et al. (2007) for example point out that too old literature may not be accurate training data for modern similarity detection tasks, because words frequently used today did not even exist back in the past.

To overcome these issues, Baroni and Lenci (2011) suggest to base the MSR scores on a standardized corpus. The one they suggest is constructed based on a combination of ukWaC (Ferraresi et al., 2008) and a 2009 Wikipedia dump, called WaCkypedia_EN, which are both available at the project page of the WaCky initiative (Baroni et al., 2009)[1].

Even though the corpus is mainly web derived, the creators made sure to meet certain quality criteria and finally evaluated ukWaC through word list comparison with the British National Corpus (BNC) (Ferraresi et al., 2008). This makes the combination of ukWaC with WaCkypedia_EN a very large general-purpose corpus of the English language and should be considered as a very good choice for experimentation regarding MSRs[2].

Another factor that greatly affects results is the definition of a "context" (see Baroni and Lenci (2011)). We experimented with different sizes and came to the conclusion that for our evaluations, two words should be in the same context if they were two sentences apart (i.e. a context is a three sentence window in a larger document). This serves as a compromise between looking at whole documents and direct term neighborhoods.

4. Performance Evaluation

There are multiple methods to evaluate the performance of MSRs. In this section we are reviewing and discussing popular ones and present the results for the chosen MSRs discussed in section 2.

4.1. Legacy Tests

One of the earliest test sets containing human judgement on semantic similarity was created by Rubenstein and Goodenough (1965) as early as 1965 during an experiment with up to 36 college students, which were to rate the "similarity of meaning" of in total 65 pairs of nouns on a scale from 0 to 4. They used the data they obtained during the experiment to conclude "that a pair of words is highly synonymous if their contexts show a relatively great amount of overlap".

Later on, Miller and Charles (1991) used a subset of 30 of the original 65 noun pairs of Rubenstein and Goodenough's test set in another experiment, in which they obtained new ratings from 0 to 4 by 38 students, which they

[1] `http://wacky.sslmit.unibo.it`

[2] Although we encourage researchers to use the corpus by Baroni et al. (2009), due to limited time and resources the test results we present in the following chapters are based on a corpus generated by a dump of the English Wikipedia.

MSR	Correlation
PMI	0.666
NSS	**0.686**
BCM	0.634

(a) Overall

MSR	Correlation
PMI	0.478
NSS	0.647
BCM	**0.662**

(b) Grouped

Table 1: Spearman correlation results of tested MSRs for the WordSim-353 dataset.

found to have an exceptionally high correlation with the original scores from 1965.

We consider these test sets as "legacy tests" not because the ratings on semantic similarity are quite old and may be outdated (which is probably not true, as Miller and Charles' results suggest), but because the original test set was not designed to provide a benchmark for MSRs, but to prove a much more general point, which we cited before. Thus, the choice of word pairs does not provide a very good coverage of different types of semantic relations we would expect for a benchmark set, not to mention the very small size of the set of only 65 word pairs.

4.2. WordSim-353

The WordSim-353 dataset (Finkelstein et al., 2001) consists of 353 word pairs rated by 16 human subjects. It includes the 65 pairs used in Rubenstein and Goodenough's set. Given a word pair, the subjects had to assign a score between zero (totally unrelated) and ten (very related). To evaluate a MSR the authors of the dataset reported correlation scores with the human judgement.

According to Baroni and Lenci (2011), WordSim-353 suffers mainly from the large variety of different types of semantic relations within the set. Evaluating a MSR favoring synonyms over meronyms may yield worse results than another one, which is not a better algorithm in general, but just favors different semantic relations.

Despite these drawbacks, we conducted two evaluations using the WordSim-353 test with our implementations of PMI, NSS, and BCM, following the procedure in Finkelstein et al. (2001). In the first test we calculated all scores for all pairs and compared the resulting rankings with the human judgements calculating the Spearman's rank correlation. This evaluation measures overall correlation. The second test groups term pairs with a common term and calculates the Spearman's rank correlation for each group. The resulting correlation values are then weighted with their group size and the average is calculated. Differences are significant comparing the distributions of group correlations using a Tukey Honestly Significance test ($\alpha = 0.05$).

The results of the test in table 1 show that all measures, maybe with the exception of PMI on the grouped test, perform reasonably well, even in the overall correlation.

4.3. Free Association Norms

In a large-scale experiment Nelson et al. (2004) asked over 6000 experiment participants to freely associate a given term with any other term that came to their minds. The given cue terms consisted of 76% nouns, 13% adjectives and 7% verbs. By aggregating the results of the subjects, they created a data source containing 72186 term pairs, that can be used as a test set for MSRs.

Although the authors suggest to use their data "for evaluating statical models of semantic representation", the experiment setup did not explicitly suggest that the participants should find semantically related words but freely associate. This leads to certain effects such as 49 out of 152 people (32.2%) responded to the cue *instinct* with *basic* (the authors explain this with the well-known 1992 movie *Basic Instinct*). However, if multiple people associated the same term with a cue, it is probable that there is a semantic relation between the two.

Lindsey et al. (2007) suggested an experiment setup that takes for each cue the set of given responses as *truly* semantically related and add a set of randomly chosen words of the same size as *false* ones. A MSR is evaluated by rating all pairs in the resulting set and ordering them by relatedness. If there are n *true* relations, they defined the accuracy as the number of *true* relations in the best n tuples divided by the number of *true* relations. For this test setup the dataset contains of 5018 groups, each containing on average 28.8 term pairs.

Maki et al. (2004) also used the *forward strength* of the norms, that is the percentage of people who chose a specific target for a given cue word, to compare correlation with the results of a MSR. With an LSA approach they reported a 0.267 correlation ($N = 49362$).

For the "discriminate-from-random" test setup in table 2, again, all MSRs are performing reasonably well, with BCM performing best when calculating the average accuracy over all groups as well as when weighting the accuracy with the size of the group. Although the differences in accuracy seem small, because of the large test set all pairwise differences are again significant ($\alpha = 0.05$).

Looking at the results of the correlation tests, we report average Spearman's rank correlation per group. BCM performs best but the correlation values are much lower than for WordSim-353. This is probably due to the fact that the "forward strength" in free associations is in itself only a measure that is merely correlated to semantic relatedness.

The most obvious drawback of this benchmark has been discussed above: The simple fact that the dataset has not been created to evaluate semantic relatedness. A minor problem is the uncontrolled random sample. Though, the set allows to test a vast amount of different types of semantic relations.

MSR	Average	Weighted
PMI	0.773	0.767
NSS	0.780	0.774
BCM	**0.796**	**0.791**

(a) "Discriminate-from-random" accuracy

MSR	Correlation
PMI	0.220
NSS	0.243
BCM	**0.283**

(b) Forward strength correlation

Table 2: Results for the tested MSRs on the Free Association Norms dataset.

4.4. TOEFL

Landauer and Dumais (1997) introduced the TOEFL synonymy detection task as a benchmark for MSRs[3]. The test includes 80 questions, each consisting of a given target term and four possible choices. One of those choices has to be identified as a synonym to the target term, the remaining answers are no synonyms of the target. The average non native english speaker in the USA achieves a mean accuracy of 64.5% (Landauer and Dumais, 1997). While at first sight it seems great to measure the capability of an algorithm to identify synonyms, this restriction to only one type of semantic relation can also be considered a drawback (Baroni and Lenci, 2011). Due to the uncoherent pattern of choices for the distractions, a MSRs general performance is hard to measure using this benchmark set. Finally, each answer being evaluated binary causes large differences in the accuracy: One more correctly answered question yields 1.25% more accuracy with only 80 test cases.

Evaluating the three MSRs on the TOEFL test set, all are displaying the same performance: 62.5% accuracy, which is 50 out of 80 correct answers. The reached score, however, is rather low compared to other published results on this popular test, which reach up to 100% accuracy (Bullinaria and Levy, 2012) with specifically tuned algorithms and parameters. Another reason for the low performance probably is that for the task of synonymy detection, it would be better not to compare the semantic similarity of terms itself but to compare the similarity of their contexts.

4.5. BLESS

Baroni and Lenci (2011) have designed their BLESS dataset specifically with the evaluation of MSRs in mind.

The main problem they see with current testing of MSRs is that researchers often "compar[e] a single quality score", such as TOEFL accuracy or WordSim-353 correlation, to evaluate if a specific MSR *outperforms* others. While we do not think that it is a generally bad idea to quantify the performance of a MSR, we strongly agree that it is important to reflect on "how and why the models differ". The BLESS set is designed specifically to give more insight into the latter.

The dataset consists of 200 nouns with 5 different *true* types of relations as well as 3 *random* types. With each of the 8 relation types containing on average about 133 related terms, the whole dataset consists of 26554 concept-type-relatum tuples. To assess a specific MSR's behavior regarding different types of semantic relations, the authors have chosen the 5 classes of semantic relations as co-hyponym or coordinate, hypernym, meronym, attribute, and event. The former three being nouns, attributes being adjectives and events verbs. The three random types are divided by their word type, with nouns, adjectives and verbs as a controlled group of non-semantically related terms to the original nouns. The authors have given great thought to the selection of the terms in their dataset, which is much more thoroughly described in Baroni and Lenci (2011).

[3]We would like to thank Dr. Thomas K. Landauer and his team at the University of Colorado for providing the TOEFL test set.

MSR	Accuracy
PMI	0.783
NSS	**0.784**
BCM	0.77

Table 3: "Discriminate-from-random" accuracy using the BLESS dataset.

The suggested experiment setup is based on box plots of per-concept z-transformed similarity scores. For each concept the authors pick the best result, *nearest neighbor*, per relation type in order to average out model specific preferences (e.g. a preference towards technical terms). The resulting 8 transformed values are then collected for all 200 concepts and presented in a box plot. Significance of comparisons is then evaluated using a Tukey Honestly Significant Difference test.

Since the authors criticized other tests to aggregate the evaluation results to only one measure of quality, such as accuracy, they don't provide guidance on how to tell if one tested MSR is *better* than another. One obviously desirable feature for the tested MSRs is discrimination of the *truly* related types from the *random* types. Thus, the test set can also be used to compare MSRs with the experiment setups used for example in the evaluation with the Free Association Norms (see section 4.3.). Even if the resulting accuracy value will certainly not state if one MSR is better or worse than another, the resulting values still are interesting to compare and interpret.

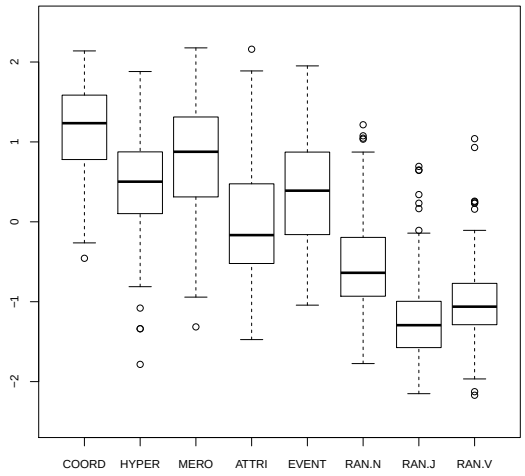
Using the set, we noted that some relata have been added very often, such as *old* (associated to 157 of 200 concepts) or *new* (96 of 200) in the attribute group. While almost everything, even abstract concepts, can be considered old or new, we think the relation is most of the times very weak. Also the word pairs sometimes rely on the supplied *part-of-speech* tagging which might lead to problems if one is not using them. However, these are minor drawbacks compared to problems other test sets show.

As displayed in figure 1, all measures are well able to discriminate the *true* relata from the random ones. Differences between each *true* relation type and the random types are all significant ($\alpha = 0.05$). All measures tend to have a distinctive preference towards co-hyponyms with BCM showing the highest preference.

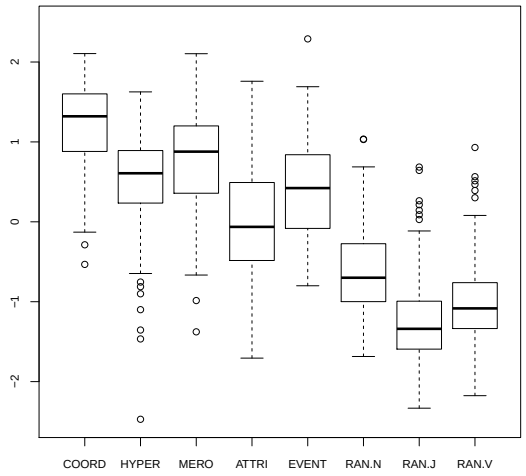
For BCM, median values for all true types but co-hyponyms are smaller than for NSS and PMI. Still, all true types but hypernyms and events are significantly differently distributed. The random types are statistically indistinguishable from each other and are packed much tighter than for NSS and PMI.

Looking at the plot and pairwise comparisons for PMI and NSS, hypernyms and events are just as in BCM indistinguishable, while interestingly both measures show, in contrast to BCM, significantly different distributions among all random group types.

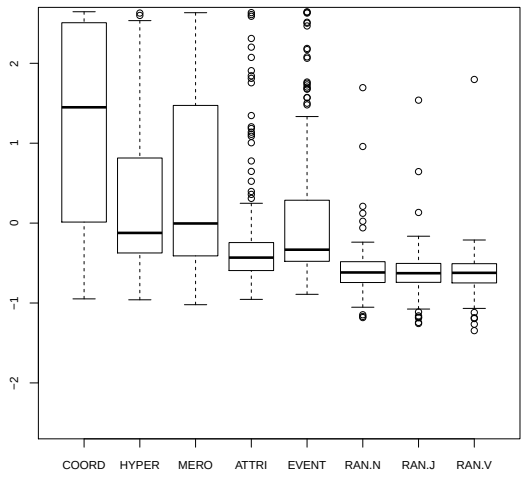
For the results of the "discriminate-from-random" test displayed in table 3, BCM performs slightly worse than PMI and NSS, but the distribution of accuracy values over all groups does not differ significantly from each other.



(a) Boxplot of PMI results



(b) Boxplot of NSS results



(c) Boxplot of BCM results

Figure 1: Distribution of z-transformed semantic similarities across concept types

5. Conclusion

In the former sections we motivated the importance, especially for information retrieval systems, of being able to determine the semantic relatedness between terms. With *PMI* and *NSS* we presented two state of the art measures and introduced the Bidirectional Co-occurrence Measure as an intuitive and well-performing alternative to the former.

We then covered the methodology of generating the underlying co-occurrence statistics and presented a survey of current benchmark tests as well as an evaluation for all three MSRs, which showed that BCM performed either on-par or significantly better than the standard measures. However, we also argued that the presented test can not definitely answer if one measure is strictly better than another. Thus, we encourage researchers and developers who are experimenting with semantic similarity to consider using BCM, but also to try different approaches to determine which specific MSRs' behavior suits their needs best.

6. Further Work

Looking at a new MSR, it would be interesting to look more into its performance when varying the underlying parameters, most prominently the used corpus and the definition of a *context window*, but also experimenting with different weighting schemes such as *BM25* instead of *idf*. Also, we would like to look into comparisons with other techniques such as ones, for example, based on ontologies or external sources like *WordNet* (Fellbaum, 1998).

References

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta, 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M. and A. Lenci, 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics.

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

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

Cilibrasi, R.L. and P.M.B. Vitanyi, 2007. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383.

El-Yaniv, R. and D. Yanay, 2012. Supervised learning of semantic relatedness. *Machine Learning and Knowledge Discovery in Databases*:744–759.

Fellbaum, Christiane, 1998. A semantic network of english: the mother of all wordnets. *Computers and the Humanities*, 32(2-3):209–220.

Ferraresi, Adriano, Eros Zanchetta, Marco Baroni, and Silvia Bernardini, 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*. ACM.

Jones, K.S., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

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

Lindsey, R., V.D. Veksler, A. Grintsvayg, and W.D. Gray, 2007. Be wary of what your computer reads: the effects of corpus selection on measuring semantic relatedness. In *8th International Conference of Cognitive Modeling, ICCM*.

Maki, W.S., L.N. McKinley, and A.G. Thompson, 2004. Semantic distance norms computed from an electronic dictionary (wordnet). *Behavior Research Methods*, 36(3):421–431.

Miller, G.A. and W.G. Charles, 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Nelson, D.L., C.L. McEvoy, and T.A. Schreiber, 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, 36(3):402–407.

Rubenstein, H. and J.B. Goodenough, 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Turney, P.D., 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*. Springer-Verlag.

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