Comparing Word Relatedness Measures Based on Google n-grams

Aminul ISLAM Evangelos MILIOS Vlado KEŠELJ
Faculty of Computer Science
Dalhousie University, Halifax, Canada
islam@cs.dal.ca, eem@cs.dal.ca, vlado@cs.dal.ca

Abstract

Estimating word relatedness is essential in natural language processing (NLP), and in many other related areas. Corpus-based word relatedness has its advantages over knowledge-based supervised measures. There are many corpus-based measures in the literature that can not be compared to each other as they use a different corpus. The purpose of this paper is to show how to evaluate different corpus-based measures of word relatedness by calculating them over a common corpus (i.e., the Google n-grams) and then assessing their performance with respect to gold standard relatedness datasets. We evaluate six of these measures as a starting point, all of which are re-implemented using the Google n-gram corpus as their only resource, by comparing their performance in five different data sets. We also show how a word relatedness measure based on a web search engine can be implemented using the Google n-gram corpus.

Keywords: Word Relatedness, Similarity, Corpus, Unsupervised, Google n-grams, Trigrams.

Proceedings of COLING 2012: Posters, pages 495–506,
COLING 2012, Mumbai, December 2012.

1 Introduction

Word relatedness between two words refers to the degree of how much one word has to do with another word whereas word similarity is a special case or a subset of word relatedness. A word relatedness method has many applications in NLP, and related areas such as information retrieval (Xu and Croft, 2000), image retrieval (Coelho et al., 2004), paraphrase recognition (Islam and Inkpen, 2008), malapropism detection and correction (Budanitsky and Hirst, 2006), word sense disambiguation (Schutze, 1998), automatic creation of thesauri (Lin, 1998a; Li, 2002), predicting user click behavior (Kaur and Hornof, 2005), building language models and natural spoken dialogue systems (Fosler-Lussier and Kuo, 2001), automatic indexing, text annotation and summarization (Lin and Hovy, 2003). Most of the approaches of determining text similarity use word similarity (Islam and Inkpen, 2008; Li et al., 2006). There are other areas where word similarity plays an important role. Gauch et al. (1999) and Gauch and Wang (1997) applied word similarity in query expansion to provide *conceptual retrieval* which ultimately increases the relevance of retrieved documents. Many approaches to spoken language understanding and spoken language systems require a grammar for parsing the input utterance to acquire its semantics. Meng and Siu (2002) used word similarity for semi-automatic grammar induction from unannotated corpora where the grammar contains both semantic and syntactic structures. An example in other areas is database schema matching (Islam et al., 2008).

Existing work on determining word relatedness is broadly categorized into three major groups: corpus-based (e.g., Cilibrasi and Vitanyi, 2007; Islam and Inkpen, 2006; Lin et al., 2003; Weeds et al., 2004; Landauer et al., 1998), knowledge-based (e.g., Radinsky et al., 2011; Gabrilovich and Markovitch, 2007; Jarmasz and Szpakowicz, 2003; Hirst and St-Onge, 1998; Resnik, 1995), and hybrid methods (e.g., Li et al., 2003; Lin, 1998b; Jiang and Conrath, 1997). Corpus-based could be either supervised (e.g., Bollegala et al., 2011) or unsupervised (e.g., Iosif and Potamianos, 2010; Islam and Inkpen, 2006). In this paper, we will focus only on unsupervised corpus-based measures.

Many unsupervised corpus-based measures of word relatedness, implemented on different corpora as resources (e.g., Islam and Inkpen, 2006; Lin et al., 2003; Weeds et al., 2004; Landauer et al., 1998; Landauer and Dumais, 1997), can be found in literature. These measures generally use co-occurrence statistics (mostly word n-grams and their frequencies) of target words generated from a corpus to form probability estimates. As the co-occurrence statistics are corpus-specific, most of the existing corpus-based measures of word relatedness implemented on different corpora are not fairly comparable to each other even on the same task. In practice, most corpora do not have readily available co-occurrence statistics usable by these measures. Again, it is very expensive to precompute co-occurrence statistics for all possible word tuples using the corpus as the word relatedness measures do not know the target words in advance. Thus, one of the main drawbacks of many corpus-based measures is that they are not feasible to be used on-line. There are other corpus-based measures that use web page count of target words from search engine as co-occurrence statistics (e.g., Iosif and Potamianos, 2010; Cilibrasi and Vitanyi, 2007; Turney, 2001). The performance of these measures are not static as the contents and the number of web pages are constantly changing. As a result, it is hard to fairly compare any new measure to these measures.

Thus, the research question arises: How can we compare a new word relatedness measure that is based on co-occurrence statistics of a corpus or a web search engine with the existing measures? We find that the use of a common corpus with co-occurrence statistics—e.g., the Google n-grams (Brants and Franz, 2006)—as the resource could be a good answer to this question. We experimentally evaluated six unsupervised corpus-based measures of word relatedness using the Google n-gram corpus on different tasks. The Google n-gram dataset[1] is a publicly available corpus with co-occurrence statistics of a large volume of web text. This will allow any new corpus based word relatedness measure to use the common corpus and compare with different existing measures on the same tasks. This will also facilitate a measure based on the Google n-gram corpus to be used on-line. Another motivation is to find an indirect mapping of co-occurrence statistics between the Google n-gram corpus and a web search engine. This is also to show that the Google n-gram corpus could be a good resource to many of the existing and future word relatedness measures. One of the previous works of this nature is (Budanitsky and Hirst, 2006), where they evaluate five knowledge-based measures of word relatedness using WordNet as their central resource.

The reasons of using corpus-based measures are threefold. First, to create, maintain and update lexical databases or resources—such as WordNet (Fellbaum, 1998) or Roget's Thesaurus (Roget, 1852)—requires significant expertise and efforts (Radinsky et al., 2011). Second, coverage of words in lexical resources is not quite enough for many NLP tasks. Third, such lexical resources are language specific, whereas Google n-gram corpora are available in English and in 10 European Languages (Brants and Franz, 2009).

The rest of this paper is organized as follows: Six corpus-based measures of word relatedness are briefly described in Section 2. Evaluation methods are discussed in Section 3. Section 4 and 5 present the experimental results from two evaluation approaches to compare several measures. We address some contributions and future related work in Conclusion.

2 Unsupervised corpus-based Approaches

Corpus-based approaches to measuring word relatedness generally use co-occurrence statistics (mostly word n-grams) of a target word from a corpus in which it occurs and then these co-occurrence statistics may be used to form probability estimates. Different corpus-based measures use different corpora to collect these co-occurrence statistics. The notation used in all the measures of word relatedness described in this section are shown in Table 1. Corpus-

Notation	Description
$C(w_1 \cdots w_n)$	frequency of the n-gram, $w_1 \cdots w_n$, where $n \in \{1, \cdots, 5\}$
$D(w_1 \cdots w_n)$	number of web documents having n-gram, $w_1 \cdots w_n$, where $n \in \{1, \cdots, 5\}$
$M(w_1, w_2)$	number of tri-grams that start with w_1 and end with w_2
$\mu_T(w_1, w_2)$	$\frac{1}{2}(\sum_{i=3}^{M(w_1,w_2)+2} C(w_1 w_i w_2) + \sum_{i=3}^{M(w_2,w_1)+2} C(w_2 w_i w_1))$, which represents the mean frequency of $M(w_1, w_2)$ tri-grams that start with w_1 and end with w_2, and $M(w_2, w_1)$ tri-grams that start with w_2 and end with w_1
N	total number of web documents used in Google n-grams
$\|V\|$	total number of uni-grams in Google n-grams
$C_{\max}$	maximum frequency possible among all Google uni-grams, i.e., $C_{\max} = \max(\{C(w_i)\}_{i=1}^{\|V\|})$

Table 1: Notation used for all the measures

[1] Details can be found at www.ldc.upenn.edu/Catalog/docs/LDC2006T13/readme.txt

based measures of word relatedness that use co-occurrence statistics directly collected from the web using a search engine (e.g., Iosif and Potamianos, 2010; Cilibrasi and Vitanyi, 2007; Turney, 2001) can not directly be implemented using the Google n-gram corpus. This is because these measures use some co-occurrence statistics which are not available in the Google n-gram corpus. Though there is no direct mapping between the Google n-gram corpus and a web search engine, it is possible to get an indirect mapping using some assumptions. It is obvious that based on the notation of Table 1, $C(w_1) \geq D(w_1)$ and $C(w_1w_2) \geq D(w_1w_2)$. This is because a uni-gram or a bi-gram may occur multiple times in a single document. Thus, considering the lower limits of $C(w_1)$ and $C(w_1w_2)$, two assumptions could be: (1) $C(w_1) \approx D(w_1)$ and (2) $C(w_1w_2) \approx D(w_1w_2)$. Based on these assumptions, we will use $C(w_1)$ and $C(w_1w_2)$ instead of using $D(w_1)$ and $D(w_1w_2)$, respectively to implement measures using the Google n-gram corpus.

2.1 Jaccard Coefficient

Jaccard coefficient (Salton and McGill, 1983) is defined as:

$$\text{Jaccard}(w_1, w_2) = \frac{D(w_1w_2)}{D(w_1) + D(w_2) - D(w_1w_2)} \approx \frac{C(w_1w_2)}{C(w_1) + C(w_2) - C(w_1w_2)} \tag{1}$$

In probability terms, Equation (1) represents the maximum likelihood estimate of the ratio of the probability of finding a web document where words w_1 and w_2 co-occur over the probability of finding a web document where either w_1 or w_2 occurs[2].

2.2 Simpson Coefficient

The Simpson coefficient is useful in minimizing the effect of unequal size of the number of web documents where the occurrence of w_1 and w_2 are mutually exclusive. Simpson or overlap coefficient (Bollegala et al., 2011) is defined as:

$$\text{Simpson}(w_1, w_2) = \frac{D(w_1w_2)}{\min(D(w_1), D(w_2))} \approx \frac{C(w_1w_2)}{\min(C(w_1), C(w_2))} \tag{2}$$

which represents the maximum likelihood estimate of the ratio of the probability of finding a web document where words w_1 and w_2 co-occur over the probability of finding a web document where the word with the lower frequency occurs.

2.3 Dice Coefficient

Dice coefficient (Smadja et al., 1996; Lin, 1998b,a) is defined as:

$$\text{Dice}(w_1, w_2) = \frac{2D(w_1w_2)}{D(w_1) + D(w_2)} \approx \frac{2C(w_1w_2)}{C(w_1) + C(w_2)} \tag{3}$$

which represents the maximum likelihood estimate of the ratio of twice the probability of finding a web document where words w_1 and w_2 co-occur over the probability of finding a web document where either w_1 or w_2 or both occurs.

[2] Normalization by the total number of web documents, N, is the same for the nominator and denominator, and can be ignored.

2.4 Pointwise Mutual Information

Pointwise Mutual Information (PMI) is a measure of how much one word tells us about the other. PMI is defined as:

$$\mathrm{PMI}(w_1, w_2) = \log_2 \left(\frac{\frac{D(w_1 w_2)}{N}}{\frac{D(w_1)}{N} \frac{D(w_2)}{N}} \right) \approx \log_2 \left(\frac{\frac{C(w_1 w_2)}{N}}{\frac{C(w_1)}{N} \frac{C(w_2)}{N}} \right) \tag{4}$$

where N is the total number of web documents. PMI between two words w_1 and w_2 compares the probability of observing the two words together (i.e., their joint probability) to the probabilities of observing w_1 and w_2 independently. PMI was first used to measure word similarity by Church and Hanks (1990). Turney (2001) used PMI, based on statistical data acquired by querying a Web search engine to measure the similarity of pairs of words.

2.5 Normalized Google Distance (NGD)

Cilibrasi and Vitanyi (2007) proposed a page-count-based distance metric between words, called the Normalized Google Distance (NGD). Normalized Google Distance relatedness between w_1 and w_2, $\mathrm{NGD}(w_1, w_2)$ is defined as:

$$\mathrm{NGD}(w_1, w_2) = \frac{\max(\log D(w_1), \log D(w_2)) - \log D(w_1 w_2)}{\log N - \min(\log D(w_1), \log D(w_2))} \tag{5}$$

$$\approx \frac{\max(\log C(w_1), \log C(w_2)) - \log C(w_1 w_2)}{\log N - \min(\log C(w_1), \log C(w_2))} \tag{6}$$

NGD is based on normalized information distance (Li et al., 2004), which is motivated by Kolmogorov complexity. The values of Equation (5) and (6) are unbounded, ranging from 0 to ∞. Gracia et al. (2006) proposed a variation of Normalized Google Distance in order to bound the similarity value in between 0 and 1, which is:

$$\mathrm{NGD}'(w_1, w_2) = e^{-2 \times \mathrm{NGD}(w_1, w_2)} \tag{7}$$

2.6 Relatedness based on Tri-grams (RT)

Islam et al. (2012) used Google n-grams, the Google tri-grams in particular, for determining the similarity of a pair of words. Their tri-gram word relatedness model can be generalized to n-gram word relatedness model. The main idea of the tri-gram relatedness model is to take into account all the tri-grams that start and end with the given pair of words and then normalize their mean frequency using uni-gram frequency of each of the words as well as the most frequent uni-gram in the corpus used. Word relatedness between w_1 and w_2 based on Tri-grams, $\mathrm{RT}(w_1, w_2) \in [0, 1]$ is defined as:

$$\mathrm{RT}(w_1, w_2) = \begin{cases} \frac{\log \frac{\mu_T(w_1, w_2) C_{\max}^2}{C(w_1) C(w_2) \min(C(w_1), C(w_2))}}{-2 \times \log \frac{\min(C(w_1), C(w_2))}{C_{\max}}} & \text{if } \frac{\mu_T(w_1, w_2) C_{\max}^2}{C(w_1) C(w_2) \min(C(w_1), C(w_2))} > 1 \\ \frac{\log 1.01}{-2 \times \log \frac{\min(C(w_1), C(w_2))}{C_{\max}}} & \text{if } \frac{\mu_T(w_1, w_2) C_{\max}^2}{C(w_1) C(w_2) \min(C(w_1), C(w_2))} \leq 1 \\ 0 & \text{if } \mu_T(w_1, w_2) = 0 \end{cases} \tag{8}$$

3 Evaluation Methods

One of the commonly accepted approaches to evaluate word relatedness measures is a comparison with human judgments. Considering human judgments of similarity or relatedness as the upper limit, this approach gives the best assessment of the 'closeness' and ‘goodness’ of a measure with respect to human judgments. Another approach is to evaluate the measures with respect to a particular application. If a system uses a measure of word relatedness (often in back end) in one of the phases, it is possible to evaluate different measure of word relatedness by finding which one the system is most effective with, while keeping all other phases of the system constant. In the remainder of this paper, we will use these two approaches to compare measures mentioned in sections 2.1 to 2.6.

4 Comparison with Human Ratings of Semantic Relatedness

4.1 Rubenstein and Goodenough’s 65 Word Pairs

Rubenstein and Goodenough (1965) conducted quantitative experiments with a group of 51 human judges who were asked to rate 65 pairs of word (English) on the scale of 0.0 to 4.0, according to their similarity of meaning. A word relatedness measure is evaluated using the correlation between the relatedness scores it produces for the word pairs in the benchmark dataset and the human ratings. The correlation coefficients of the six implemented measures with the human judges for the 65 word pairs from Rubenstein and Goodenough (1965) dataset (henceforth, R&G dataset) are shown in Figure 1.

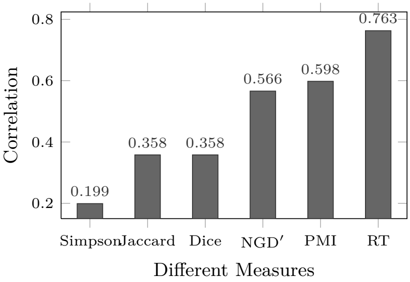

Figure 1: Similarity correlations on RG’s 65 noun pairs.

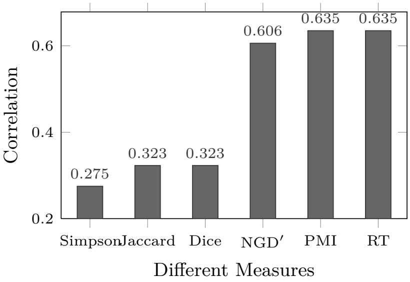

Figure 2: Similarity correlations on MC’s 28 noun pairs.

4.2 Miller and Charles’ 28 Noun Pairs

Miller and Charles (1991) repeated the same experiment (done by Rubenstein and Goodenough, 1965) restricting themselves to 30 pairs from the original 65, and then obtained similarity judgments from 38 human judges. Most researchers used 28 word pairs of the Miller and Charles (1991) dataset (henceforth, M&C dataset), because two word pairs were omitted from the earlier version of WordNet. The correlation coefficient of different measures with the human judges for 28 word pairs from M&C dataset are shown in Figure 2. It is shown in Figure 2 that the correlation coefficients for both PMI and RT on M&C dataset are same, whereas Figure 1 shows RT’s improvement of 16.5 percentage points over PMI on R&G dataset.

5 Application-based Evaluation of Measures of Relatedness

5.1 TOEFL's 80 Synonym Questions

Consider the following synonym test question which is one of the 80 TOEFL (Test of English as a Foreign Language) questions from Landauer and Dumais (1997): Given the problem word *infinite* and the four alternative words *limitless*, *relative*, *unusual*, *structural*, the task is to choose the alternative word which is most similar in meaning to the problem word. The number of correct answers for different word relatedness measures on 80 TOEFL questions is shown in Figure 3. RT measure gets 65 per cent correct answers. A human average score on the same question set is 64.5 per cent (Landauer and Dumais, 1997).

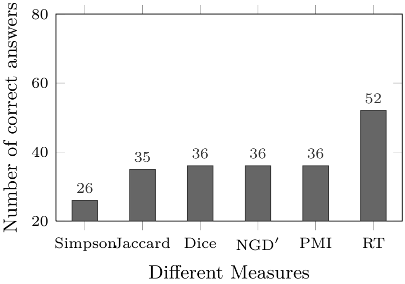

Figure 3: Results on TOEFL's 80 synonym questions.

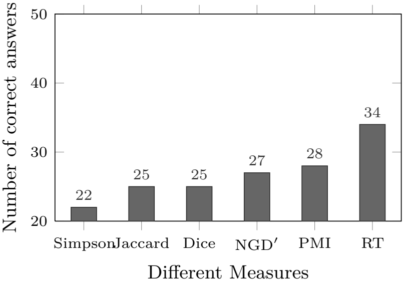

Figure 4: Results on ESL's 50 synonym questions.

5.2 ESL's 50 Synonym Questions

The task here is the same as TOEFL's 80 synonym questions task, except that the synonym questions are from the English as a Second Language (ESL) tests. The number of correct answers for different measures on 50 ESL synonym questions is shown in Figure 4.

5.3 Text Similarity

The task of text similarity is to find the similarity between two text items. The idea is to use all the discussed word relatedness measures separately on a single text similarity measure and then evaluate the results of the text similarity measure based on a standard data set used for the task to see which word relatedness measure works better. There are many text similarity measures, both supervised and unsupervised, in the literature that use word similarity in the back end (e.g., Li et al., 2006; Liu et al., 2007; Feng et al., 2008; O'Shea et al., 2008; Islam and Inkpen, 2008; Ho et al., 2010; Tsatsaronis et al., 2010; Islam et al., 2012). We use one of the state-of-the-art unsupervised text similarity measures proposed by Islam et al. (2012) to evaluate all the discussed word relatedness measures. One of the reasons of using this text similarity measure is that it only uses the relatedness scores of different word pairs in the back end. The main idea of the text similarity measure proposed by Islam et al. (2012) is to find for each word in the shorter text, some most similar matchings at the word level, in the longer text, and then aggregate their similarity scores and normalize the result.

In order to evaluate the text similarity measure, we compute the similarity score for 30 sentence pairs from Li et al. (2006) and find the correlation with human judges. The details of this data set preparation are in (Li et al., 2006). This is one of the most used data sets for evaluating the task. For example, Li et al. (2006); Liu et al. (2007); Feng et al. (2008); O'Shea et al. (2008); Islam and Inkpen (2008); Ho et al. (2010); Tsatsaronis et al. (2010); Islam et al. (2012) used the same 30 sentence pairs and computed the correlation with human judges. The correlation coefficients of Islam et al. (2012) text similarity measures (based on the discussed word relatedness measures) with the human judges for 30 sentence pairs are shown in Figure 5. On the 30 sentence pairs, Ho et al. (2010) used one of the state-of-the-art word relatedness measures using WordNet to determine the relatedness scores of word pairs, then applied those scores in Islam and Inkpen (2008) text similarity measure, and achieved a Pearson correlation coefficient of 0.895 with the mean human similarity ratings. On the same dataset, Tsatsaronis et al. (2010) achieved a Pearson correlation coefficient of 0.856 with the mean human similarity ratings. Islam et al. (2012) text similarity measure using RT achieves a high Pearson correlation coefficient of 0.916 with the mean human similarity ratings which is close to that of the best human participant. The improvement achieved over Ho et al. (2010) is statistically significant at 0.05 level.

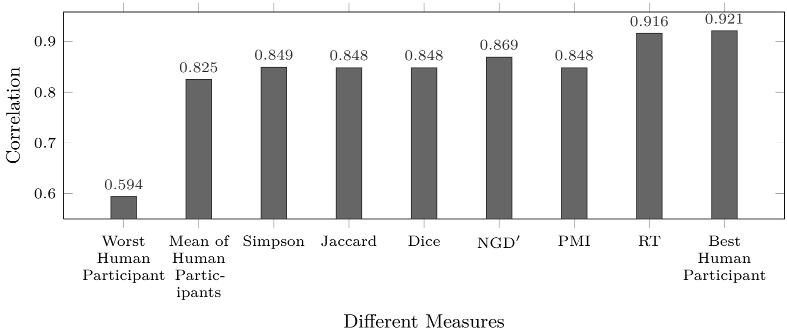

Figure 5: Similarity correlations on Li's 30 sentence pairs.

Conclusion

This paper shows that any new corpus-based measure of word relatedness that uses n-gram statistics can easily be implemented on the Google n-gram corpus and be *fairly* evaluated with existing works on standard data sets of different tasks. We also show how to find an indirect mapping of co-occurrence statistics between the Google n-gram corpus and a web search engine using some assumptions. One of the advantages of measures based on n-gram statistics is that they are language independent. Although English is the focus of this paper, none of the word relatedness measures discussed in this paper depends on any specific language, and could be used with almost no change with many other languages that have a sufficiently large n-gram corpus available. Future work could be to evaluate other corpus-based measures using the common Google n-gram corpus and the standard data sets for different tasks.

References

Bollegala, D., Matsuo, Y., and Ishizuka, M. (2011). A web search engine-based approach to measure semantic similarity between words. *IEEE Trans. on Knowl. and Data Eng.*, 23(7):977–990.

Brants, T. and Franz, A. (2006). Web 1T 5-gram corpus version 1.1. Technical report, Google Research.

Brants, T. and Franz, A. (2009). Web 1T 5-gram, 10 European languages version 1. Technical report, Linguistic Data Consortium, Philadelphia.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

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

Coelho, T., Calado, P., Souza, L., Ribeiro-Neto, B., and Muntz, R. (2004). Image retrieval using multiple evidence ranking. *IEEE Transaction on Knowledge and Data Engineering*, 16(4):408–417.

Fellbaum, C., editor (1998). *WordNet: An electronic lexical database.* MIT Press.

Feng, J., Zhou, Y.-M., and Martin, T. (2008). Sentence similarity based on relevance. In Magdalena, L., Ojeda-Aciego, M., and Verdegay, J., editors, *IPMU*, pages 832–839.

Fosler-Lussier, E. and Kuo, H.-K. (2001). Using semantic class information for rapid development of language models within ASR dialogue systems. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:553–556.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gauch, S. and Wang, J. (1997). A corpus analysis approach for automatic query expansion. In *Proceedings of the sixth international conference on Information and knowledge management*, CIKM '97, pages 278–284, New York, NY, USA. ACM.

Gauch, S., Wang, J., and Rachakonda, S. M. (1999). A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Trans. Inf. Syst.*, 17(3):250–269.

Gracia, J., Trillo, R., Espinoza, M., and Mena, E. (2006). Querying the web: a multiontology disambiguation method. In *Proceedings of the 6th International Conference on Web Engineering*, ICWE '06, pages 241–248, New York, NY, USA. ACM.

Hirst, G. and St-Onge, D. (1998). *WordNet: An electronic lexical database*, chapter Lexical chains as representations of context for the detection and correction of malapropisms, pages 305–332. The MIT Press, Cambridge, MA.

Ho, C., Murad, M. A. A., Kadir, R. A., and Doraisamy, S. C. (2010). Word sense disambiguation-based sentence similarity. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 418–426, Stroudsburg, PA, USA. Association for Computational Linguistics.

Iosif, E. and Potamianos, A. (2010). Unsupervised semantic similarity computation between terms using web documents. *IEEE Trans. on Knowl. and Data Eng.*, 22(11):1637–1647.

Islam, A. and Inkpen, D. (2006). Second order co-occurrence PMI for determining the semantic similarity of words. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1033–1038, Genoa, Italy.

Islam, A. and Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data*, 2:10:1–10:25.

Islam, A., Inkpen, D., and Kiringa, I. (2008). Applications of corpus-based semantic similarity and word segmentation to database schema matching. *The VLDB Journal*, 17(5):1293–1320.

Islam, A., Milios, E. E., and Keselj, V. (2012). Text similarity using google tri-grams. In Kosseim, L. and Inkpen, D., editors, *Canadian Conference on AI*, volume 7310 of *Lecture Notes in Computer Science*, pages 312–317. Springer.

Jarmasz, M. and Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *RANLP*, volume 260 of *Current Issues in Linguistic Theory (CILT)*, pages 111–120. John Benjamins, Amsterdam/Philadelphia.

Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33.

Kaur, I. and Hornof, A. J. (2005). A comparison of LSA, WordNet and PMI-IR for predicting user click behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 51–60, New York, NY, USA. ACM.

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

Landauer, T., Foltz, P., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Li, H. (2002). Word clustering and disambiguation based on co-occurrence data. *Nat. Lang. Eng.*, 8(1):25–42.

Li, M., Chen, X., Li, X., Ma, B., and Vitanyi, P. M. B. (2004). The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264.

Li, Y., Bandar, Z., and Mclean, D. (2003). An approach for measuring semantic similarity using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. on Knowl. and Data Eng.*, 18:1138–1150.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 71–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.

Lin, D. (1998b). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003). Identifying synonyms among distributionally similar words. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 1492–1493, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Liu, X., Zhou, Y., and Zheng, R. (2007). Sentence similarity based on dynamic time warping. In *Proceedings of the International Conference on Semantic Computing*, pages 250–256, Washington, DC, USA. IEEE Computer Society.

Meng, H. H. and Siu, K. C. (2002). Semiautomatic acquisition of semantic structures for understanding domain-specific natural language queries. *IEEE Trans. on Knowl. and Data Eng.*, 14(1):172–181.

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

O'Shea, J., Bandar, Z., Crockett, K., and McLean, D. (2008). A comparative study of two short text semantic similarity measures. In *Proceedings of the 2nd KES International Conference on Agent and Multi-Agent Systems: Technologies and Applications*, KES-AMSTA'08, pages 172–181, Berlin, Heidelberg. Springer-Verlag.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA. ACM.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Roget, P. (1852). *Roget's Thesaurus of English Words and Phrases.* Penguin Books; 150th Anniversary edition (July 2007).

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

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill.

Schutze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Comput. Linguist.*, 22(1):1–38.

Tsatsaronis, G., Varlamis, I., and Vazirgiannis, M. (2010). Text relatedness based on a word thesaurus. *J. Artif. Int. Res.*, 37(1):1–40.

Turney, P. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML 2001)*, pages 491–502, Freiburg, Germany.

Weeds, J., Weir, D., and McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xu, J. and Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112.