Using Filtered Second Order Co-occurrence Matrix to Improve the Traditional Co-occurrence Model

Chun-Hung Lu[a*], Chorng-Shyong Ong[a] ，Wen-Lian Hsu[b,] Hsing-Kuo Lee[b]
[a]Department of Information Technology Management, National Taiwan University,
#1, Sec 4, Roosecelt Road, Taipei 10617, Taiwan(R.O.C)
[b]Institute of Information Science, Academia Sinica,
#128 Academia Road Section 2 Nankang, Taipei, Taiwan115
*Corresponding Author: d95006@im.ntu.edu.tw

ABSTRACT

Using co-occurrence statistics to measure word similarities/relatedness has applications in many areas of natural language processing. Our experiment results also indicate that two words with zero co-occurrence statistics could still be related. In this paper, we present two algorithms, both of which were evaluated on 80 synonym test questions from the Test of English as a Foreign Language (TOEFL) and 50 synonym test questions from a collection of tests for students of English as a Second Language (ESL). The evaluation results show that the first algorithm improves the performance of co-occurrence based applications significantly; and the second ensemble algorithm (which incorporates the first algorithm) achieves the best results on the synonym questions of both tests.

Keyword: Co-occurrence, Word similarity, Word relatedness, Synonym test, PMI

1. Introduction

Many statistical tests have been proposed to measure the strength of word similarities or word associations in natural language texts (Dunning, 1993; Church and Hanks, 1990; Dagan et al., 1999). Such tests attempt to measure the dependency between words by using statistics taken from a large corpus. In this context, a key assumption is that similarity between words is a consequence of word co-occurrence, or that the proximity of the words in the text is indicative of some kind of relationship between them, such as synonymy or antonymy.

Generally, linguists use two types of methods to improve raw frequency counts: asymptotic hypothesis tests and mutual information. The two approaches are conceptually different and typically produce rather different types of results. All of the hypothesis tests and mutual information methods start by calculating how many times one might expect to find a word pair in a corpus of a certain size simply by chance, given the frequencies of the component words.

However, the above hypothesis is quite restrictive, especially in terms of similarity measurement, because many synonymous words may not appear in the same sentence or within a certain distance (window size) of one another. French & Labiouse (2002) posited that co-occurrence statistics alone are not sufficient to build complete and reliable lexical representations because the co-occurrence statistics cannot deal with homophones and homographs. Our experiment results also indicate that two words with zero co-occurrence statistics could still be related. Based on our experiment results, we designed two algorithms to prove that two words with zero co-occurrence statistics could still be related.

The first algorithm tries to use the information in the co-occurrence model to enhance the similarity measure. By using the same test corpus (ESL & TOEFL questions) and open source software, we provide evidence that filtered second order co-occurrence matrix derived from the co-occurrence matrix can improve the co-occurrence statistics significantly. Based on these results and some other tests reported in the literature, our second algorithm employs an ensemble strategy to measure word similarities. The algorithm outperforms all existing methods.

The contribution of this paper is two-fold. First, we introduce and evaluate a method that enhances the results of co-occurrence-related algorithms for answering multiple-choice synonym questions. The algorithm may be useful for solving problems in lexical semantics, such as determining semantic orientations, and semantic relations. Second, we present a novel product rule for combining module outputs. Compared to exist approaches, it achieves the best results on ESL (English as a Second language) and TOEFL (Test of English as a Foreign Language) tests.

The remainder of this paper is organized as follows. In Section 2, we review the literature on Pointwise Mutual Information (PMI). In Section 3, we present our algorithms and discuss our experiment results. Then, in Section 4, we summarize our findings and indicate future research directions.

2. Related Works

2.1 Applications of co-occurrence statistics

Co-occurrence statistics play an important role in the field of natural-language processing. Co-occurrences represent the observable evidence that can be distilled from a corpus by fully automatic means. After statistical generalization, the information can be used to predict which word combinations are most likely to appear in another corpus. In addition to these direct uses, co-occurrence data often serve as the basis for distributional methods, which compare the “co-occurrence profile” of a given word, a vector of association scores for its co-occurrences, with the profiles of other words. The distance between two such vectors (which can be defined in various ways) is interpreted as an indicator of their semantic similarity. Clustering and dimensionality reduction methods, such as factor analysis or singular-value decomposition, can then be used to identify classes of semantically related words. The following are some applications of distributional techniques:

- detecting semantic similarities between words (Landauer and Dumais 1997), especially synonyms (Turney 2001; Rapp 2002; Terra and Clarke 2003);
- unsupervised induction of word senses, usually combined with disambiguation of the senses identified automatically (Pantel and Lin 2002; Rapp 2003);
- identification of translation equivalents (which are semantically related, of course) in non-parallel corpora, i.e., unrelated texts in two or more languages (Rapp 1999);
- distinguishing between compositional and lexicalized compound nouns, based on the assumption that the former are more similar to their head noun (Zinsmeister and Heid 2004);
- selecting informative clauses for the compilation of biographical summaries (Schiffman et al. 2001);
- classifying all types of texts or adjectives according to their sentiments (Turney and Littman 2003);
- conducting computational experiments with the Appraisal framework, and assigning adjectives to one of the three broad attitude classes (Taboada and Grieve 2004);
- using PMI as a feature in relation extraction (Pasca et al. 2006)

Semantic similarity is a central concept in many fields, such as artificial intelligence, natural language processing, cognitive science and psychology. A number of methods have been proposed for measuring the similarity between two words. In this paper, we use synonyms tests to demonstrate that the proposed algorithms can improve the results of previous works based on co-occurrence statistics. First, we review the PMI method in the next sub-section.

2.2 The literature on PMI

Pointwise Mutual Information (PMI), first used in the context of word association by Church and Hanks (1990), is a very simple information-theoretic measure. When computed between two words W*i* and W*j*, PMI “compares the probability of observing W*i* and W*j* together (the joint probability) with the probabilities of observing W*i* and W*j* independently (chance)” (Church & Hanks, 1990, p. 23). It is defined as follows:

$$\mathrm{PMI}(W_i, W_j) = \log_2 \frac{p(W_i \& W_j)}{p(W_i) * p(W_j)}$$

In practice, P(W$_{i}$) can be approximated as the number of times that W*i* appears in the target corpus; P(W$_{j}$) as the number of times y appears in the corpus; and P(W*i* and W*j*) as the number of times the two words co-occur in a context.

Landauer and Dumais (1997) used Latent Semantic Analysis (LSA), a word similarity measure, to answer TOEFL (Test of English as a Foreign Language) synonym questions. Their approach achieved 64.37% accuracy on the questions. Subsequently, Turney (2001) proposed using PMI to analyze data collected by information retrieval (PMI-IR) as an unsupervised measure for evaluating the semantic similarity of words. The method uses PMI and the web corpus of the AltaVista search engine to estimate word similarities. He proposed four scoring methods to measure co-occurrences and used synonyms tests to compare the results obtained on AltaVista’s corpus of 350 million web pages with results obtained by LSA run on an encyclopedia of 30,473 articles. PMI’s performance was comparable to, or better than, that of LSA on TOEFL. The PMI-IR score is calculated as follows:

$$\mathrm{PMI-IR}(W_i, W_j) = \log_2 \frac{p(W_i \& W_j)}{p(W_i) * p(W_j)}$$

The formula, which indicates the degree of statistical dependence between W*i* (the problem) and W*j* (the choice), can be used as a measure of the semantic similarity of W*i* and W*j*. Of the four types of queries suggested by Turney (2001), the following query used to collect counts from the AltaVista search engine achieved the best result (PMI_IR score3):

$$p_{score}(\text{problem \&choice}) = \frac{\text{hits(problem NEAR choice}_i\text{)AND NOT((problem OR choice) NEAR not))}}{\text{hits(choice}_i\text{ AND NOT (choice}_i\text{ NEAR not))}}$$

In a set of experiments based on TOEFL synonymy tests (Turney 2001), the PMI-IR measure using the NEAR operator identified the correct answer (out of four synonym choices) in 72.5% of the cases.

Terra and Clarke (2003) conducted a comparative investigation of co-occurrence frequency estimation on the performance of synonym tests. They reported that PMI (with a certain window size) outperformed, the compared methods with an average of 81.25% of the questions answered correctly.

Ruiz-Casado et al. (2005) proposed a method, called *context overlapping*, for obtaining synonyms from a large corpus or the Web. It determines the semantic similarity of two terms, and achieved 82.5% accuracy on TOEFL's synonym test.

Generalized Latent Semantic Analysis (GLSA) (Matveeva et al., 2005) is a technique that combines the strengths of PMI and LSA. Like LSA, it uses dimensionality reduction (SVD) to filter out noise in the system. However, instead of adopting LSA's initial word-by-document co-occurrence matrix, it utilizes a word-by-word PMI matrix in which words are represented as vectors of PMI scores relative to other words in the vocabulary (Niwa & Nitta, 1994).

Budiu, Royer, and Pirolli (2007) found that PMI outperformed LSA on a variety of semantic tasks when trained on a larger corpus, but their work confounded corpus size for corpus quality. Bullinaria and Levy, using PMI to compare vectors of co-occurrence counts (i.e., the rows of a term * term matrix), found that PMI outperformed LSA on the TOEFL forced-choice synonymy test when both methods were trained on a small corpus derived from Grolier's Academic American Encyclopedia. However, they did not correlate the performance of PMI and LSA with human judgments of semantic similarity (Recchia and M. N. Jones, 2009).

3. Algorithm Design and Experiment

3.1 Test Data

Synonymy tests are often used to test students' vocabulary knowledge. A question in a synonymy test presents a word and asks the examinee to choose the word that is most similar to it from among several (usually four) alternatives. The performance on this test is computed as the percentage of questions answered correctly out of the total number of questions. We used two synonymy tests: Test of English as a Foreign Language (TOEFL), first used by Landauer & Dumais(1997); and English as a Second Language (ESL) (Turney, 2001). Both tests assess the language proficiency of foreign students.

TOEFL. TOEFL is comprised of 80 questions drawn from test materials designed by the Educational Testing Service to measure English proficiency of foreign students who wish to continue their education in the US. The test was first used by Landauer and Dumais (1997) to compare the results obtained by LSA with those of students. The items in the 80 questions are all single words; a sample question is “Select the synonym of the word *flawed* from the set containing imperfect, tiny, lustrous, and crude”. The authors reported that foreign college students typically score around 65% on TOEFL’s synonymy questions.

ESL. The ESL test, which contains 57 questions about single words, which first used by Turney (2001) to compare the performance of PMI and LSA. Sample items include *rusty* (choices: corroded, black, dirty, painted) and *lump* (choices: chunk, stem, trunk, limb).

3.2 Algorithm I. Improving PMI-based computing with a filtered second order co-occurrence matrix

To obtain co-occurrence statistics, researchers usually modify a model with a wordlist and record the co-occurrence statistics for two words W$_i$ and W$_j$ that belong to the wordlist.

Recall that two words with zero co-occurrence statistics could still be related. To demonstrate the point, we trained PMI on a corpus derived from Wikipedia. The so-called “Wikipedia corpus” was comprised of 4.1 GB of text in English Wikipedia articles downloaded in 2008. We used WP2TXT (http://wp2txt.rubyforge.org/) to extract plain text data from the Wikipedia dump file (encoded in XML/compressed with Bzip2) and stripped all the MediaWiki markups and other metadata.

To ensure the experiment was fair, we used an open source tool, called Lightweight Metrics of Semantic Similarity (LMOSS, http://mypage.iu.edu/~grecchia/lmoss.html),

PMI filtered second order score 1

$$= \frac{1}{n} * \sum_{1\sim n} p(W_i, W_n) * p(W_j, W_n) + \frac{p(W_i \& W_j)}{p(W_i) * p(W_j)}$$

$$\cong \frac{1}{n} * \sum_{1\sim n}\left(\left(\frac{p(W_i \& W_n)}{\sqrt{\frac{p(W_i)}{C}}}\right) * \left(\frac{* P(W_n \& W_j)}{\sqrt{\frac{p(W_j)}{C}}}\right)\right) + \frac{p(W_i \& W_j)}{\sqrt{\frac{p(W_i)}{C}} * \sqrt{\frac{p(W_j)}{C}}}$$

Formula 1

PMI filtered second order score 2

$$\frac{\#\text{ of words overlaid in N}}{N} + \frac{p(W_i \& W_j)}{\sqrt{\frac{p(W_i)}{C}} * \sqrt{\frac{p(W_j)}{C}}}$$

to train the model and present the results. We also obtained some general wordlists from the National Puzzlers' League and Brian Kelk’s website. To evaluate LMOSS on the wordlists, we modified some of the tool’s parameters and ran it on a 64-bit 2 *Core*™2 Quad HP ProLiant DL 360 G5 with 14 GB of memory and a processor speed of 1.6 GHz. The results, listed in Table 1, show that 50,000~100,000 words achieve a better performance on the 4G Wikipedia corpus. A larger wordlist would probably cause more ambiguity.

We analyzed the results and found that some related words did not co-occur in a particular window size. For example, we rarely use “*buy*” and “*purchase*” in a same sentence. This phenomenon also occurs in the concept of event frames (Fillmore, 1982) in traditional linguistics. Fillmore’s concept of frame semantics characterizes the semantic and syntactic properties of predicating words by relating them to semantic frames. This means that similar or related concepts will appear in a certain context, but not necessarily within a certain window size. In addition, event frames provide some evidence that related words can co-occur with the same word in the same text. By representing this concept as a graph, we can use the concept edge between W_i ,W_n and the concept edge between W_j and W_n to compute the relation between W_i and W_j (as shown in Fig. 1. (b)). Therefore, we modified the PMI model as shown in Formula 1.

Table 2 shows the results of PMI filtered second order score1. Although the results are better than those of the original PMI model, computing the score required a substantial amount of time (4 minutes for the ESL Test). Hence, we considered a simplified formula (Formula 2). In Formula 2, given N most frequent words in the co-occurrence wordlist matrix of W_i and W_j respectively, we compared the overlaid words between these two sets. Table 3 shows the results of PMI filtered second orderscore2 using 110,000 words with the 4G Wikipedia corpus, whose performance is closer to the result of PMI-IR score2.

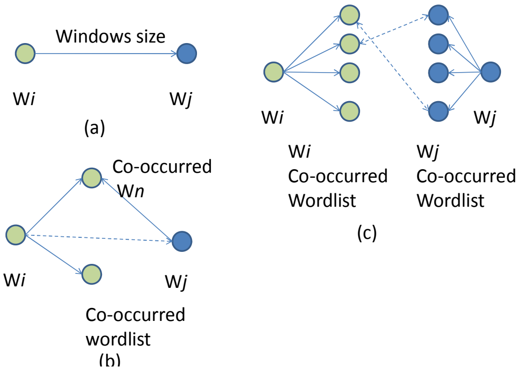

Figure 1 Representation of the relation between W_i and W_j

Comparing the results in Tables 1, 2, 3, the PMI filtered second order score2 achieved the best results on different sized wordlists with significance levels of more than 5%. The results tend to support our hypothesis that two words with zero co-occurrence statistics can still be related.

The above findings motivated us to incorporate PMI filtered second order score2 into PMI-IR method. The “web as corpus” approach provides a huge corpus and an almost infinite wordlist. However, it also has drawbacks in terms of performance because most search engines have strict limits on the size and frequency of queries, which render the PMI-IR method impractical. For example, the PMI-IR formula given in Section 2.2 requires 600 queries in the ESL test and PMI-IR score3 requires 400 queries. After implementing the PMI-IR algorithm and assessing its performance, we found that, even if we ignore the search engine’s restrictions, PMI-IR required 14 minutes and 23 seconds to complete the ESL test. The response times of different

Table 1 The results of different wordlists trained on LMOSS with Wikipedia text by traditional PMI

Wordlist	BNC	Wikipedia Most common words (Gutenberg)	UK English wordlist (Brain Klek)	CMU dictionary	Merriam-Webster's 9th Collegiate	Webster's New International Dictionary
Words	6318	36663	57046	111308	120996	234936
TOEFL Test	6	24	43	45	34	31
ESL Test	5	15	19	19	18	17

Table 2 The results of different wordlists trained on LMOSS with Wikipedia text by applying PMI reference score1

Wordlist	BNC	Wikipedia Most common words (Gutenberg)	UK English wordlist (Brain Klek)	CMU dictionary	Merriam-Webster's 9th Collegiate	Webster's New International Dictionary
TOEFL Test	10	23	48	53	31	31
ESL Test	12	21	19	20	19	20

Table 3 The results of different wordlists trained on LMOSS with Wikipedia text by applying PMI reference score2

Wordlist	BNC	Wikipedia Most common words (Gutenberg)	UK English wordlist (Brain Klek)	CMU dictionary	Merriam-Webster's 9th Collegiate	Webster's New International Dictionary
TOEFL Test	16	37	51	56*	44	39
ESL Test	13	22	20	22	22	22

methods are detailed in Table 4. To implement our PMI filtered second order score2 algorithm in PMI-IR, we modified the co-occurrence wordlist matrix by counting the word information obtained from snippets of 100 search results. By removing stop words and comparing the overlay of the co-occurrence wordlist of W$_i$ and W$_j$, we improved the PMI-IR method’s performance for both tests, as shown by the results in Table 5.

Table 4 The response times of different methods

Methods /Tests	Original PMI	PMI with reference Score2	PMI-IR Score2	PMI-IR Score3
ESL Test	1 second	15 seconds	10 minutes and 12 seconds	14 minutes and 23 seconds

Table 5 The PMI-IR cross reference scores

Methods/Tests	PMI-IR Score2	PMI-IR Score3	PMI-IR with reference Score2
TOEFL Test	57	64	65
ESL Test	20	31	32

Table 6. The results of different independent methods and merging rules on synonym test

				Ensemble Methods		
Methods	Second-String (JaroWinkler)	Wordnet (Definition & Synonyms)	PMI-IR reference score2	Our Ensemble result	Jamasz & Szpakowicz , 2003	Turney et al., 2003
TOEFL Test	23 (28.75%)	44 (55%)	65 (81.8%)	78 (97.5%)	N/A	78 (97.5%,)
ESL Test	13 (26%)	27 (54%)	32 (64%)	43 (86%)	41 (82%,)	N/A

$$\begin{aligned} &\text{Similarity Score} \\ &= C_i * \text{Similarity}_{\text{Edit distance}} + C_j * \text{Similarity}_{\text{Dictionary}} + C_k * \text{Similarity}_{\text{Frequencies}} \\ &\text{where } C_i + C_j + C_k = 1 \end{aligned}$$

Formula 3

3.3 Ensemble method

In our empirical study, we combined three independent methods. To reduce the possibility of bias in our approach, we used open-source tools based on Formula 3. In the formula, C_i, C_j, C_K represents a constant that C_i,+ C_j,+ C_K = 1 in order to ensemble the result.

3.3.1 Edit Distance based Similarity Measure

We use SecondString, an open-source Java toolkit for name-matching methods, as a distance-based similarity measure. The distance function is a fundamental component of SecondString. Usually, a distance functions maps a pair of strings to a real number, where a smaller value of the real number indicates greater similarity between the strings. In our study, we calculate the synonym similarity score by adapting the Jaro distance metric.

3.3.2 Dictionary-based Similarity Measure

There are several dictionary-based approaches for measuring the similarity of words. Most of them use WordNet, a broad coverage lexical network of English words, but some use Roget’s Thesaurus. We use the WordNet.Net API developed by Michael Crowe and Troy Simpson; however, we extend the word similarity function, which is implemented in their module, by adding the condition “if W_j exists in W_i’s synonyms.

3.3.3 Statistical Frequency-based Similarity Measure

We use the PMI-IR filtered second order model as our Statistical Frequency-based method.

3.4 Experiment

Table 6 presents the results of testing the four modules (edit distance, dictionary, frequencies and ensemble method) on synonym problems. The PMI-IR filtered second order score achieved the highest accuracy among the individual methods (80%). The ensemble method, ESL and TOEFL, achieved 86% and 97.5 %(TOEFL) respectively. Statistically, the results are significantly better than those of the best individual module in this study, as well the results of previous studies. It seems that this domain lends itself to ensemble approaches. Our ensemble method achieved the highest accuracy on the ESL and TOEFL tests, listed on the ACL website (http://www.aclweb.org/aclwiki/index.php?title=ESL_Synonym_Questions_(State_of_the_art , http://www.aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions).

4.Conclusion and Future work

In this paper, we have presented an algorithm that improves the performance of the co-occurrence statistics-based model in similarity testing. The algorithm provides an easy-to-use and highly accurate resource for obtaining similarity judgments from corpora. In addition, we combined different methods and achieved the best results on ESL/TOEFL synonym test.

In our future work, we will utilize a larger wordlist compiled from the Web to test the PMI-IR filtered second order model and investigate whether referencing by wordlists (with more stable high frequency words and a larger N) could improve the performance of the co-occurrence model significantly. We will also apply our method to different languages.

REFERENCES

Aminul Islam , Diana Inkpen , Iluju Kiringa (2008). Applications of corpus-based semantic similarity and word segmentation to database schema matching, The VLDB Journal - The International Journal on Very Large Data Bases, v.17 n.5, p.1293-1320, August 2008

Budiu, R., Royer, C., & Pirolli, P. L. (2007). Modeling information scent: a comparison of lsa, pmi and glsa similarity measures on common tests and corpora. In 8th riao conference. Pittsburgh, PA.

Bullinaria, J.A., and Levy, J.P. (2006). Extracting semantic representations from word co-occurrence statistics: A computational study. To appear in Behavior Research Methods, 38.

Charles J. Fillmore. (1982). Frame semantics. In Linguistics in the Morning Calm, pages 111~137.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information and lexicography. Computational Linguistics, 16, 22-29.

Donna Harman. 1992. Relevance feedback revisited. In Proceedings of 1992 SIGIR conference, Copenhagen, Denmark.

Hirst, G., and St-Onge, D. (1998). Lexical chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum (ed.), WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 305-332.

Dagan, I., Lee, L., Pereira, F. (1999). Similarity-based models of word co-occurrence probabilities. Machine Learning 34(1-3):43-69.

French, R. M., & Labiouse, C. (2002). Four problems with extracting human semantics from large text corpora. Proceedings of the Twenty- Fourth Annual

Conference of the Cognitive Science Society (pp. 316- 322). Mahwah, NJ: Erlbaum.

Jarmasz, M., and Szpakowicz, S. (2003). Roget’s thesaurus and semantic similarity, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03), Borovets, Bulgaria, September, pp. 212-219.

Jiang, J.J., and Conrath, D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the International Conference on Research in Computational Linguistics, Taiwan.

Kenneth Ward Church and Patrick Hanks. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22–29.

Landauer, T.K., and Dumais, S.T. (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.

Leacock, C., and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (ed.), WordNet: An Electronic Lexical Database. Cambridge: MIT Press, pp. 265-283.

Lin, D. (1998). An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning (ICML-98), Madison, WI, pp. 296-304.

Maite Taboada and Jack Grieve.(2004). Analyzing Appraisal automatically. In Spring Symposium on Exploring Attitude and Affect in Text. American Association for Artificial Intelligence, Stanford. AAAI Technical Report SS-04-07

Matveeva, I., Levow, G., Farahat, A., and Royer, C. (2005). Generalized latent semantic analysis for term representation. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria.

Pasca M., Lin D., Bigham J., Lifchits A., and Jain A.. (2006). Names and similarities on the web: Fact extraction in the fast lane. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics(ACL 2006), pp. 809–816.

Pado, S., and Lapata, M. (2007). Dependency-based construction of semantic space models. Computational Linguistics, 33(2), 161-199.

Pantel, Patrick and Lin, Dekang (2002). Discovering word senses from text. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 613–619, Edmonton, Canada.

Rapp, Reinhard (1999). Automatic identification of word translations from unrelated English and German corpora. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics(ACL 1999), Maryland.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. Proceedings of the Ninth Machine Translation Summit, pp. 315-322.

Resnik, P. (1995). Using information content to evaluate semantic similarity. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, pp. 448-453.

Ruiz-Casado, M., Alfonseca, E. and Castells, P. (2005) Using context-window overlapping in Synonym Discovery and Ontology Extension. Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP-2005), Borovets, Bulgaria.

Recchia and M. N. Jones. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information to latent semantic analysis. Behavior Research Methods.

Schiffman, Barry; Mani, Inderjeet; Concepcion, Kristian J. (2001). Producing biographical summaries: Combining linguistic knowledge with corpus statistics. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics(ACL 2001), . pp.458-465

Ted Dunning.(1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19:61–74.

Terra, E., and Clarke, C.L.A. (2003). Frequency estimates for statistical word similarity measures. Proceedings of the Human Language Technology and North American Chapter of Association of Computational Linguistics Conference 2003 (HLT/NAACL 2003), pp. 244–251.

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

Turney, P.D.. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), pp. 417-424.

Turney, P.D., Littman, M.L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03), Borovets, Bulgaria, pp. 482-489.

Turney, P.D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, pp. 905-912.

W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string metrics for matching names and records. In Proceedings of the workshop on Data Cleaning and Object Consolidation at the International Conference on Knowledge Discovery and Data Mining (KDD), 2003.

Zinsmeister, Heike and Heid, Ulrich (2004). Collocations of complex nouns: Evidence for lexicalisation. In Proceedings of the 11th Euralex International Congress, Lorient,France.

Wordlist Source

Brian Kelk’s website, http://www.bckelk.ukfsn.org/menu.html

http://www.puzzlers.org/dokuwiki/doku.php?id=solving:wordlists:about:start

Wiktionary:Frequency lists, http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists