Frequency Estimates for Statistical Word Similarity Measures

Egidio Terra
School of Computer Science
University of Waterloo
elterra@math.uwaterloo.ca

C. L. A. Clarke
School of Computer Science
University of Waterloo
claclark@plg2.uwaterloo.ca

Abstract

Statistical measures of word similarity have application in many areas of natural language processing, such as language modeling and information retrieval. We report a comparative study of two methods for estimating word co-occurrence frequencies required by word similarity measures. Our frequency estimates are generated from a terabyte-sized corpus of Web data, and we study the impact of corpus size on the effectiveness of the measures. We base the evaluation on one TOEFL question set and two practice questions sets, each consisting of a number of multiple choice questions seeking the best synonym for a given target word. For two question sets, a context for the target word is provided, and we examine a number of word similarity measures that exploit this context. Our best combination of similarity measure and frequency estimation method answers 6-8% more questions than the best results previously reported for the same question sets.

1 Introduction

Many different statistical tests have been proposed to measure the strength of *word similarity* or *word association* in natural language texts (Dunning, 1993; Church and Hanks, 1990; Dagan et al., 1999). These tests attempt to measure dependence 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 closeness of the words in text is indicative of some kind of relationship between them, such as synonymy or antonymy.

Although word sequences in natural language are unlikely to be independent, these statistical tests provide quantitative information that can be used to compare pairs of co-occurring words. Also, despite the fact that word co-occurrence is a simple idea, there are a variety of ways to estimate word co-occurrence frequencies from text. Two words can appear close to each other in the same document, passage, paragraph, sentence or fixed-size window. The boundaries for determining co-occurrence will affect the estimates and as a consequence the word similarity measures.

Statistical word similarity measures play an important role in information retrieval and in many other natural language applications, such as the automatic creation of thesauri (Grefenstette, 1993; Li and Abe, 1998; Lin, 1998) and word sense disambiguation (Yarowsky, 1992; Li and Abe, 1998). Pantel and Lin (2002) use word similarity to create groups of related words, in order to discover word senses directly from text. Recently, Tan et al. (2002) provide an analysis on different measures of independence in the context of association rules.

Word similarity is also used in language modeling applications. Rosenfeld (1996) uses word similarity as a constraint in a maximum entropy model which reduces the perplexity on a test set by 23%. Brown et al. (1992) use a word similarity measure for language modeling in an interpolated model, grouping similar words into classes. Dagan et al. (1999) use word similarity to assign probabilities to unseen bigrams by using similar bigrams, which reduces perplexity up to 20% in held out data.

In information retrieval, word similarity can be used to identify terms for pseudo-relevance feedback (Harman, 1992; Buckley et al., 1995; Xu and Croft, 2000; Vechtomova and Robertson, 2000). Xu and Croft (2000) expand queries under a pseudo-relevance feedback model by using similar words from documents retrieved and improve effectiveness by more than 20% on an 11-point average precision.

Landauer and Dumais (1997) applied word similarity measures to answer TOEFL (Test Of English as a Foreign Language) synonym questions using Latent Semantic Analysis. Turney (2001) performed an evaluation of a specific word similarity measure using the same TOEFL questions and compared the results with those obtained

C = "The results of the test were quite [unambiguous]."
TW = 'unambiguous'
A = {'clear','doubtful','surprising','illegal'}

Figure 1: Finding the best synonym option in presence of context

TW = 'boast'
A = {'brag','yell','complain','explain'}

Figure 2: Finding the best synonym

by Landauer and Dumais.

In our investigation of frequency estimates for word similarity measures, we compare the results of several different measures and frequency estimates to solve human-oriented language tests. Our investigation is based in part on the questions used by Landauer and Dumais, and by Turney. An example of such tests is the determination of the best synonym in a set of alternatives $A = \{A_1, A_2, A_3, A_4\}$ for a specific target word TW in a context $C = \{w'_1, w'_2, .., w'_n\} \setminus TW$, as shown in figure 1. Ideally, the context can provide support to choose best alternative for each question. We also investigate questions where no context is available, as shown in figure 2. These questions provides an easy way to assess the performance of measures and the co-occurrence frequency estimation methods used to compute them.

Although word similarity has been used in many different applications, to the best of our knowledge, ours is the first comparative investigation of the impact of co-occurrence frequency estimation on the performance of word similarity measures. In this paper, we provide a comprehensive study of some of the most widely used similarity measures with frequency estimates taken from a terabyte-sized corpus of Web data, both in the presence of context and not. In addition, we investigate frequency estimates for co-occurrence that are based both on documents and on a variety of different window sizes, and examine the impact of the corpus size on the frequency estimates. In questions where context is available, we also investigate the effect of adding more words from context.

The remainder of this paper is organized as follows: In section 2 we briefly introduce some of the most commonly used methods for measuring word similarity. In section 3 we present methods to assess word co-occurrence frequencies. Section 4 presents our experimental evaluation, which is followed by a discussion of the results in section 5.

2 Measuring Word Similarity

The notion for co-occurrence of two words can depicted by a *contingency table*, as shown in table 1. Each dimension represents a random discrete variable W_i with range $\mathcal{A} = \{w_i, \neg w_i\}$(presence or absence of word i in a given text window or document). Each cell in the table represent the joint frequency $f_{w_i,w_j} = N_{max} * P(i,j)$, where N_{max} is the maximum number of co-occurrences. Under an independence assumption, the values of the cells in the contingency table are calculated using the probabilities in table 2. The methods described below perform different measures of how distant observed values are from expected values under an independence assumption. Tan et al. (2002) indicate that the difference between the methods arise from non-uniform marginals and how the methods react to this non-uniformity.

	w_1	$\neg w_1$	
w_2	f_{w_1,w_2}	$f_{\neg w_1,w_2}$	f_{w_2}
$\neg w_2$	$f_{w_1,\neg w_2}$	$f_{\neg w_1,\neg w_2}$	$f_{\neg w_2}$
	f_{w_1}	$f_{\neg w_1}$	N

Table 1: Contingency table

$P(w_1, w_2) = P(w_1) * P(w_2)$
$P(\neg w_1, w_2) = P(\neg w_1) * P(w_2)$
$P(w_1, \neg w_2) = P(w_1) * P(\neg w_2)$
$P(\neg w_1, \neg w_2) = P(\neg w_1) * P(\neg w_2)$

Table 2: Probabilities under independence

Occasionally, a context C is available and can provide support for the co-occurrence and alternative methods can be used to exploit this context. The procedures to estimate $P(w_1, w_2)$, as well $P(w_i)$, will be described in section 3.

2.1 Similarity between two words

We first present methods to measure the similarity between two words w_1 and w_2 when no context is available.

2.1.1 Pointwise Mutual Information

This measure for word similarity was first used in this context by Church and Hanks (1990). The measure is given by equation 1 and is called Pointwise Mutual Information. It is a straightforward transformation of the independence assumption (on a specific point), $P(w_1, w_2) = P(w_1) * P(w_2)$, into a ratio. Positive values indicate that words occur together more than would be expected under an independence assumption. Negative values indicate

that one word tends to appear only when the other does not. Values close to zero indicate independence.

$$PMI(W_1 = w_1, W_2 = w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \quad (1)$$

2.1.2 χ^2-test

This test is directly derived from observed and expected values in the contingency tables.

$$\chi^2 = \sum_{x \in W_1} \sum_{y \in W_2} \frac{(f_{x,y} - E_{x,y})^2}{E_{x,y}} \quad (2)$$

The χ^2 statistic determines a specific way to calculate the difference between values expected under independence and observed ones, as depicted in equation 2. The values $f_{x,y}$ correspond to the observed frequency estimates.

2.1.3 Likelihood ratio

The likelihood ratio test provides an alternative to check two simple hypotheses based on parameters of a distribution. Dunning (1993) used a likelihood ratio to test word similarity under the assumption that the words in text have a binomial distribution.

Two hypotheses used are: H1:$P(w_2|w_1) = P(w_2|\neg w_1)$ (i.e. they occur independently); and H2: $P(w_2|w_1) \neq P(w_2|\neg w_1)$ (i.e. not independent). These two conditionals are used as sample in the likelihood function $L(P(w_2|w_1), P(w_2|\neg w_1); \theta)$, where θ in this particular case represents the parameter of the binomial distribution $b(n, k; \theta)$. Under hypothesis H1, $P(w_2|w_1) = P(w_2|\neg w_1) = p$, and for H2, $P(w_2|w_1) = p_1, P(w_2|\neg w_1) = p_2$.

$$\lambda = \frac{L(P(w_2|w_1); p) * L(P(w_2|\neg w_1); p)}{L(P(w_2|w_1); p_1) * L(P(w_2|\neg w_1); p_2)} \quad (3)$$

Equation 3 represents the likelihood ratio. Asymptotically, $-2log\lambda$ is χ^2 distributed.

2.1.4 Average Mutual Information

This measure corresponds to the expected value of two random variables using the same equation as PMI. Average mutual information was used as a word similarity measure by Rosenfeld (1996) and is given by equation 4.

$$MI(W_1; W_2) = \sum_{x \in W_1} \sum_{y \in W_2} P(x, y) log \frac{P(x, y)}{P(x)P(y)} \quad (4)$$

2.2 Context supported similarity

Similarity between two words can also be inferred from a context (if given). Given a context $C = \{w'_1, w'_2, ..., w'_n\}$, w_1 and w_2 are related if their co-occurrence with words in context are similar.

2.2.1 Cosine of Pointwise Mutual Information

The PMI between each context word w' and w_i form a vector. The elements in the vector represents the similarity weights of w' and w_i. The cosine value between the two vectors corresponding to w_1 and w_2 represents the similarity between the two words in the specified context, as depicted in equation 5.

$$CP(w_1; w_2) = \frac{\sum_{w' in C} PMI(w', w_1) PMI(w', w_2)}{\sqrt{\sum_{w'} PMI(w', w_1)^2} \sqrt{\sum_{w'} PMI(w', w_2)^2}} \quad (5)$$

Values closer to one indicate more similarity whereas values close to zero represent less similarity. Lesk (1969) was one of the first to apply the cosine measure to word similarity, but did not use pointwise mutual information to compute the weights. Pantel (2002) used the cosine of pointwise mutual information to uncover word sense from text.

2.2.2 L_1 norm

In this method the conditional probability of each word w'_i in C given w_1 (and w_2) is computed. The accumulated distance between the conditionals for all words in context represents the similarity between the two words, as shown in equation 6. This method was proposed as an alternative word similarity measure in language modeling to overcome zero-frequency problems of bigrams (Dagan et al., 1999).

$$L(w_1; w_2) = \sum_{w' in C} |P(w'|w1) - P(w'|w2)| \quad (6)$$

In this measure, a smaller value indicates a greater similarity.

2.2.3 Contextual Average Mutual Information

The conditional probabilities between each word in the context and the two words w_1 and w_2 are used to calculate the mutual information of the conditionals (equation 7). This method was also used in Dagan et. al. (1999).

$$AMIC(w_1; w_2) = \sum_{w'} P(w'|w1) log \frac{P(w'|w1)}{P(w'|w2)} \quad (7)$$

2.2.4 Contextual Jensen-Shannon Divergence

This is an alternative to the Mutual Information formula (equation 8). It helps to avoid zero frequency problem by averaging the two distributions and also provides a symmetric measure (AMIC is not symmetric). This method was also used in Dagan et. al. (1999).

$$KL(p\|q) = \sum p \, log \frac{p}{q}$$

$$AVGP = \frac{P(w'|w_1) + P(w'|w_2)}{2}$$

$$IRAD(w_1; w_2) = KL(P(w'|w1)\|AVGP) + KL(P(w'|w2)\|AVGP) \quad (8)$$

2.2.5 Pointwise Mutual Information of Multiple words

Turney (2001) proposes a different formula for Pointwise Mutual Information when context is available, as depicted in equation 9. The context is represented by C', which is any subset of the context C. In fact, Turney argued that bigger C' sets are worse because they narrow the estimate and as consequence can be affected by noise. As a consequence, Turney used only one word c_i from the context, discarding the remaining words. The chosen word was the one that has biggest pointwise information with w_1. Moreover, w_1 (TW) is fixed when the method is used to find the best A_i for TW, so $P(w_1, C')$ is also fixed and can be ignored, which transforms the equation into the conditional $P(w_1|w_2, C)$.

It is interesting to note that the equation $P(w_1|w_2, C)$ is not the traditional n-gram model since no ordering is imposed on the words and also due to the fact that the words in this formula can be separated from one another by other words.

$$PMIC(w_1, w_2; C') = \frac{P(w_1, w_2, C')}{P(w_2, C')P(w_1, C')} \quad (9)$$

2.2.6 Other measures of word similarities

Many other measures for word similarities exists. Tan et al. (2002) present a comparative study with 21 different measures. Lillian (2001) proposes a new word similarity measure in the context of language modeling, performing an comparative evaluation with other 7 similarity measures.

3 Co-occurrence Estimates

We now discuss some alternatives to estimate word co-occurrence frequencies from an available corpus. All probabilities mentioned in previous section can be estimated from these frequencies. We describe two different approaches: a window-oriented approach and a document-oriented approach.

3.1 Window-oriented approach

Let f_{w_i} be the frequency of w_i and the co-occurrence frequency of w_1 and w_2 be denoted by f_{w_1,w_2}. Let N be the size of the corpus in words. In the window-oriented approach, individual word frequencies are the corpus frequencies. The maximum likelihood estimate (MLE) for w_i in the corpus is $P(w_i) = f_{w_i}/N$.

The joint frequency f_{w_1,w_2} is estimated by the number of windows where the two words co-occur. The window size may vary, Church and Hanks (1990) used windows of size 2 and 5. Brown et al. (1992) used windows containing 1001 words. Dunning (1993) also used windows of size 2, which corresponds to word bigrams. Let the number of windows of size t in the corpus be N_{wt}. Recall that N_{max} is the maximum number of co-occurrences, i.e. $N_{max} = N_{wt}$ in the windows-oriented approach. The MLE of the co-occurrence probability is given by $P(w_1, w_2) = f_{w_1,w_2}/N_{wt}$.

In most common case, windows are overlapping, and in this case $N_{wt} = N - t + 1$. The total frequency of windows for co-occurrence should be adjusted to reflect the multiple counts of the same co-occurrence. One method to account for overlap is to divide the total count of windows by $window\_size - 1$. This method also reinforces closer co-occurrences by assigning them a larger weight.

Smoothing techniques can be applied to address the zero-frequency problem, or alternatively, the window size can be increased, which also increases the chance of co-occurrence. To avoid inconsistency, windows do not to cross document boundaries.

3.2 Document-oriented approach

In information retrieval, one commonly uses document statistics rather than individual word statistics. In an document-oriented approach, the frequency of a word w_i is denoted by df_{w_i} and corresponds to the number of documents in which the word appears, regardless of how frequently it occurs in each document. The number of documents is denoted by D. The MLE for an individual word in document oriented approach is $P(w_i) = df_{w_i}/D$.

The co-occurrence frequency of two words w_1 and w_2, denoted by df_{w_1,w_2}, is the number of documents where the words co-occur. If we require only that the words co-occur in the same document, no distinction is made between distantly occurring words and adjacent words. This distortion can be reduced by imposing a maximal distance for co-occurrence, (i.e. a fixed-sized window), but the frequency will still be the number of documents where the two words co-occur within this distance. The MLE for the co-occurrence in this approach is $P(w_1, w_2) = df_{w_1,w_2}/D$, since $N_{max} = D$ in the document-oriented approach.

3.3 Syntax based approach

An alternative to the Window and Document-oriented approach is to use syntactical information (Grefenstette, 1993). For this purpose, a Parser or Part-Of-Speech tagger must be applied to the text and only the interesting pairs of words in correct syntactical categories used. In this case, the fixed window can be superseded by the result of the syntax analysis or tagging process and the frequency of the pairs can be used directly. Alternatively, the number of documents that contain the pair can also be used. However, the nature of the language tests in this work make it impractical to be applied. First, the alternatives are not in a context, and as such can have more than one part-of-speech tag. Occasionally, it is possible to infer that the syntactic category of the alternatives from context of the target word TW, if there is such a context . When the alternatives, or the target word TW, are multiwords then the problem is harder, as depicted in the first example of figure 7. Also, both parsers and POS tagger make mistakes, thus introducing error. Finally, the size of the corpus used and its nature intensify the parser/POS taggers problems.

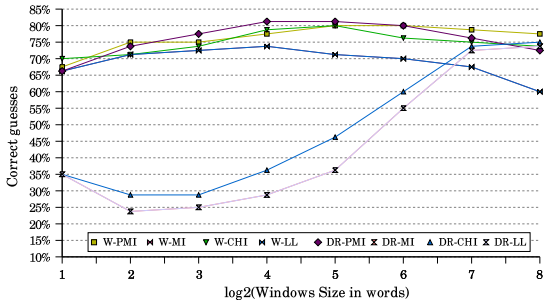

Figure 3: Results for TOEFL test set

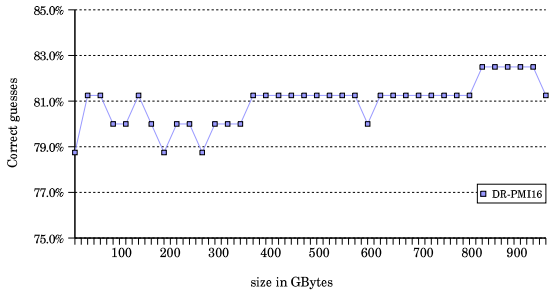

Figure 4: Impact of corpus size on TOEFL test set

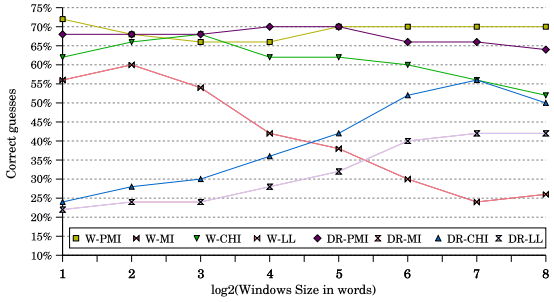

Figure 5: Results for TS1 and no context

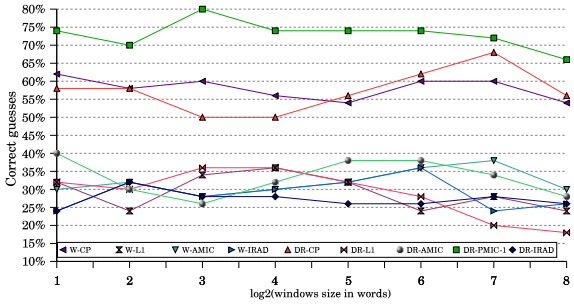

Figure 6: Results for TS1 and context

4 Experiments

We evaluate the methods and frequency estimates using 3 test sets. The first test set is a set of TOEFL questions first used by Landauer and Dumais (1997) and also by Turney (2001). This test set contains 80 synonym questions and for each question one TW and four alternative options ($|A| = 4$) are given. The other two test sets, which we will refer to as TS1 and TS2, are practice questions for the TOEFL. These two test sets also contain four alternatives options, $|A| = 4$, and TW is given in context C (within a sentence). TS1 has 50 questions and was also used by Turney (2001). TS2 has 60 questions extracted from a TOEFL practice guide (King and Stanley, 1989). For all test sets the answer to each question is known and unique. For comparison purposes, we also use TS1 and TS2 with no context.

For the three test sets, TOEFL, TS1 and TS2 without context, we applied the word and document-oriented frequency estimates presented. We investigated a variety of window sizes, varying the window size from 2 to 256 by powers of 2.

The labels used in figures 3, 5, 6, 8, 9, 10, 12 are composed from a keyword indicating the frequency estimate used (W-window oriented; and DR-document retrieval oriented) and a keyword indicating the word similarity measure. For no-context measures the keywords are: PMI-Pointwise Mutual Information; CHI-Chi-Squared; MI-Average mutual information; and LL-Log-likelihood. For the measures with context: CP-Cosine pointwise mutual information; L1-L1 norm; AMIC-Average Mutual Information in the presence of context; IRAD-Jensen-Shannon Divergence; and PMIC-n - Pointwise Mutual Information with n words of context.

For TS1 and TS2 with context, we also investigate Turney's hypothesis that the outcome of adding more words from C is negative, using DR-PMIC. The result of this experiment is shown in figures 10 and 12 for TS1 and TS2 respectively.

It is important to note that in some of the questions, TW or one or more of the A_i's are multi-word strings. For these questions, we assume that the strings may be treated as collocations and use them "as is", adjusting the size of the windows by the collocation size when applicable.

The corpus used for the experiments is a terabyte of Web data crawled from the general web in 2001. In order to balance the contents of the corpus, a breadth-first order search was used from a initial seed set of URLs representing the home page of 2392 universities and other educational organizations (Clarke et al., 2002). No duplicate pages are included in the collection and the crawler also did not allow a large number of pages from the same site to be downloaded simultaneously. Overall, the collection contains 53 billion words and 77 million documents.

A key characteristic of this corpus is that it consists of HTML files. These files have a focus on the presentation, and not necessarily on the style of writing. Parsing or tagging these files can be a hard process and prone to introduction of error in rates bigger than traditional corpora used in NLP or Information Retrieval.

We also investigate the impact of the collection size on

C = “The country is plagued by [turmoil].”
A = {‘constant change’,‘utter confusion’,‘bad weather’,‘fuel shortages’}

C = “[For] all their protestations, they heeded the judge’s ruling.”
A = {‘In spite of’,‘Because of’,‘On behalf of’,‘without’}

Figure 7: Examples of harder questions in TS2

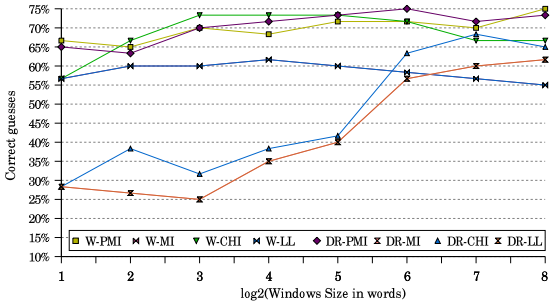

Figure 8: Results for TS2 and no context

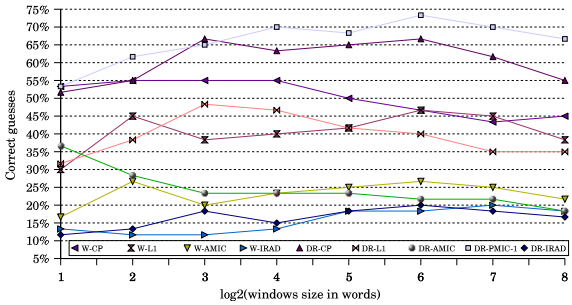

Figure 9: Results for TS2 and context

the results, as depicted in figures 4, 11 and 13 for TOEFL, TS1 and TS2 test sets, respectively.

5 Results and Discussion

The results for the TOEFL questions are presented in figure 3. The best performance found is 81.25% of the questions correctly answered. That result used DR-PMI with a window size of 16-32 words. This is an improvement over the results presented by Landauer and Dumais (1997) using Latent Semantic Analysis, where 64.5% of the questions were answered correctly, and Turney (2001), using pointwise mutual information and document retrieval, where the best result was 73.75%.

Although we use a similar method (DR-PMI), the difference between the results presented here and Turney’s results may be due to differences in the corpora and differences in the queries. Turney uses Altavista and we used our own crawl of web data. We can not compare the collections since we do not know how Altavista collection is created. As for the queries, we have more control over the queries since we can precisely specify the window size and we also do not know how queries are evaluated in Altavista.

PMI performs best overall, regardless of estimates used (DR or W). W-CHI performs up to 80% when using window estimates, outperforming DR-CHI. MI and LL yield exactly the same results (and the same ranking of the alternatives), which suggests that the binomial distribution is a good approximation for word occurrence in text.

The results for MI and PMI indicate that, for the two discrete random variables W_1 and W_2 (and range $\mathcal{A} = \{w_i, \neg w_i\}$), no further gain is achieved by calculating the expectation in the divergence. Recall that the divergence formula has an embedded expectation to be calculated between the joint probability of these two random variables and their independence. The peak of information is exactly where both words co-occur, i.e. when $W_1 = w_1$ and $W_2 = w_2$, and not any of the other three possible combinations.

Similar trends are seen when using TS1 and no context, as depicted in figure 5. PMI is best overall, and DR-PMI and W-PMI outperform each other with different windows sizes. W-CHI has good performance in small windows sizes. MI and LL yield identical (poor) results, being worst than chance for some window sizes. Turney (2001) also uses this test set without context, achieving 66% peak performance compared with our best performance of 72% (DR-PMI).

In the test set TS2 with no context, the trend seen between TOEFL and TS1 is repeated, as shown in figure 8. PMI is best overall but W-CHI performs better than PMI in three cases. DR-CHI performs poorly for small windows sizes. MI and LL also perform poorly in comparison with PMI. The peak performance is 75%, using DR-PMI with a window size of 64.

The result are not what we expected when context is used in TS1 and TS2. In TS1, figure 6, only one of the measures, DR-PMIC-1, outperforms the results from non-context measures, having a peak of 80% correct answers. The condition for the best result (one word from context and a window size of 8) is similar to the one used for the best score reported by Turney. L1, AMIC and IRAD perform poorly, worst than chance for some window sizes. One difference in the results is that for DR-PMIC-1 only the best word from context was used, while the other methods used all words but stopwords. We examine the context and discovered that using more words degrades the performance of DR-PMIC in all different windows sizes but, even using all words except stopwords, the result from DR-PMIC is better than any other contextual measure - 76% correct answers in TS1 (with DR-PMIC and a window size of 8).

For TS2, no measure using context was able to perform

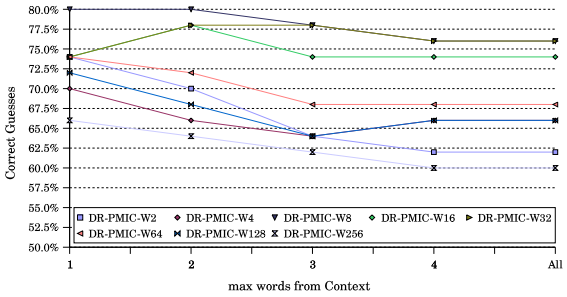

Figure 10: Influence from the context on TS1

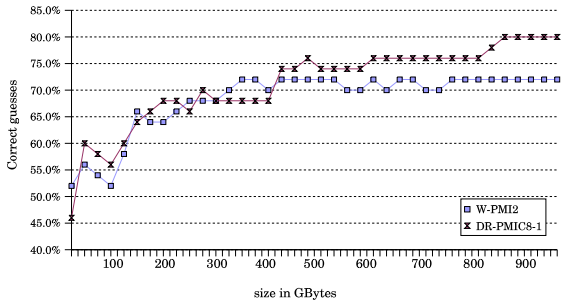

Figure 11: Impact of corpus size on TS1

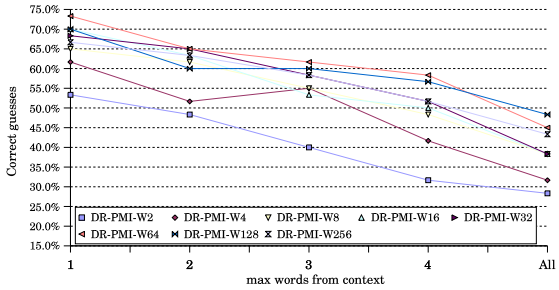

Figure 12: Influence from the context on TS2

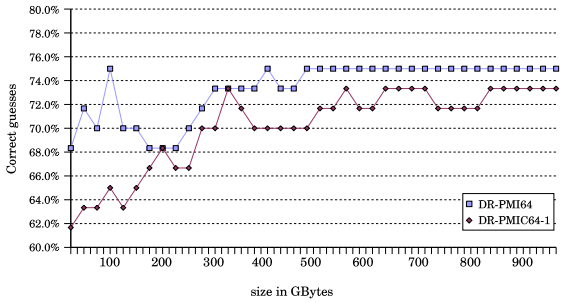

Figure 13: Impact of corpus size on TS2

better than the non-contextual measures. DR-PMIC-1 performs better overall but has worse performance than DR-CP with a window size of 8. In this test set, the performance of DR-CP is better than W-CP. L1 performs better than AMIC but both have poor results, IRAD is never better than chance. The context in TS2 has more words than TS1 but the questions seem to be harder, as shown in figure 7. In some of the TS2 questions, the target word or one of the alternatives uses functional words. We also investigate the influence of more words from context in TS2, as depicted in figure 12, where the trends seen with TS1 are repeated.

The results in TS1 and TS2 suggest that the available context is not very useful or that it is not being used properly.

Finally, we selected the method that yields the best performance for each test set to analyze the impact of the corpus size on performance, as shown in figures 4, 11 and 13. For TS1 we use W-PMI with a window size of 2 (W-PMI2) when no context is used and DR-PMIC-1 with a window size of 8 (DR-PMIC8-1) when context is used. For those measures, very little improvement is noticed after 500 GBytes for DR-PMIC8-1, roughly half of the collection size. No apparent improvement is achieved after 300-400 GBytes for W-PMI2. For TS2 we use DR-PMI with a window size of 64 (DR-PMI64) when no context is used, and DR-PMIC-1 with a windows size of 64 (DR-PMIC64-1) when context is used. It is clear that for TS2 no substantial improvement in DR-PMI64 and DR-PMIC64-1 is achieved by increasing the corpus size to values bigger than 300-400 GBytes. The most interesting impact of corpus size was on TOEFL test set using DR-PMI with a window size of 16 (DR-PMI16). Using the full corpus is no better than using 5% of the corpus, and the best result, 82.5% correct answers, is achieved when using 85-95% of corpus size.

6 Conclusion

Using a large corpus and human-oriented tests we describe a comprehensive study of word similarity measures and co-occurrence estimates, including variants on corpus size. Without any parameter training, we were able to correctly answer at least 75% questions in all test sets. From all combinations of estimates and measures, document retrieval with a maximum window of 16 words and pointwise mutual information performs best on average in the three test sets used. However, both document or windows-oriented approach for frequency estimates produce similar results in average. The impact of the corpus size is not very conclusive, it suggests that the increase in the corpus size normally reaches an asymptote, but the points where this occurs is distinct among different measures and frequency estimates.

Our results outperform the previously reported results on test sets when no context is used, being able to correctly answer 81.25% of TOEFL synonym questions, compared with a previous best result of 73.5%. A human average score on the same type of questions is 64.5% (Landauer and Dumais, 1997). We also perform better than previous work on another test set used as practice questions for TOEFL, obtaining 80% correct answers

compared to a best result of 74% from previous work.

Acknowledgments

This work was made possible also in part by PUC/RS and Ministry of Education of Brazil through CAPES agency.

References

P. F. Brown, P. V. deSouza, R. L. Mercer, T. J. Watson, V. J. Della Pietra, and J. C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

C. Buckley, G. Salton, J. Allan, and A. Singhal. 1995. Automatic query expansion using smart: Trec 3. In *The third Text REtrieval Conference*, Gaithersburg, MD.

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

C.L.A. Clarke, G.V. Cormack, M. Laszlo, T.R. Lynam, and E.L. Terra. 2002. The impact of corpus size on question answering performance. In *Proceedings of 2002 SIGIR conference*, Tampere, Finland.

I. Dagan, L. Lee, and F. C. N. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

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

G. Grefenstette. 1993. Automatic theasurus generation from raw text using knowledge-poor techniques. In *Making sense of Words. 9th Annual Conference of the UW Centre for the New OED and text Research*.

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

C. King and N. Stanley. 1989. *Building Skills for the TOEFL*. Thomas Nelson and Sons Ltd, second edition.

T. K. Landauer and S. T. Dumais. 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.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001*, pages 65–72.

M. E. Lesk. 1969. Word-word associations in document retrieval systems. *American Documentation*, 20(1):27–38, January.

Hang Li and Naoki Abe. 1998. Word clustering and disambiguation based on co-occurence data. In *COLING-ACL*, pages 749–755.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619.

R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling. computer speech and language. *Computer Speech and Language*, 10:187–228.

P.-N. Tan, V. Kumar, and J. Srivastava. 2002. Selecting the right interestingness measure for association patterns. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 32–41.

P. D. Turney. 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.

O. Vechtomova and S. Robertson. 2000. Integration of collocation statistics into the probabilistic retrieval model. In *22nd Annual Colloquium on Information Retrieval Research*, Cambridge, England.

J. Xu and B. Croft. 2000. Improving the effectiveness of information retrieval. *ACM Transactions on Information Systems*, 18(1):79–112.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes, France, July.