A Freely Available Automatically Generated Thesaurus of Related Words

Reinhard Rapp

University of Mainz, FASK
76711 Germersheim, Germany
rapp@mail.fask.uni-mainz.de

Abstract

A freely available English thesaurus of related words is presented that has been automatically compiled by analyzing the distributional similarities of words in the British National Corpus. The quality of the results has been evaluated by comparison with human judgments as obtained from non-native and native speakers of English who were asked to provide rankings of word similarities. According to this measure, the results generated by our system are better than the judgments of the non-native speakers and come close to the native speakers' performance. An advantage of our approach is that it does not require syntactic parsing and therefore can be more easily adapted to other languages. As an example, a similar thesaurus for German has already been completed.

1. Introduction

As has been shown by Ruge (1992), Grefenstette (1994), Schütze (1997), Lin (1998), Rapp (2002), and others, the semantic similarity of two words can be computed by determining the agreement of their lexical neighborhoods. For example, the semantic similarity of the words *truck* and *lorry* can be derived from the fact that they both frequently co-occur with words like *drive*, *transport*, *load*, *street*, *traffic*, *petrol*, and so forth. If for each word in a corpus a co-occurrence vector is determined whose entries are the common occurrences with all other words in the corpus, then the semantic similarities between words can be computed by conducting simple vector comparisons. To determine the words most similar to a given word, its co-occurrence vector is compared to the co-occurrence vectors of all other words in the vocabulary using one of the standard vector similarity measures; for example, the cosine coefficient. Those words that obtain the best scores are considered to be most similar.

Many researchers have used this type of context analysis as a basis to determine semantically related words. However, to improve results, several modifications of the basic algorithm have been suggested. For example, Ruge (1992), Grefenstette (1994), and Lin (1998) perform a syntactic analysis beforehand and only look at word pairs that are in a certain relation to each other, e.g. a head-modifier, verb-object, or subject-object relation. In contrast, Landauer & Dumais (1997) found that applying a *singular value decomposition* (SVD) on the underlying co-occurrence matrix improves results. And Sahlgren (2001) uses *random indexing* for better efficiency.

In this paper we present an optimized algorithm, adopt an evaluation method known as the TOEFL synonym test, and compare the accuracy of our results with the accuracies reported in related publications. We also provide human synonym judgments as obtained in an experiment conducted with native and non-native speakers of English.

2. TOEFL synonym test

Given many different proposals for the computation of semantically related words, an evaluation of the different algorithms is desirable. Although most researchers have compared their results to some gold standard, unfortunately the resources used as the standard have been widely different, ranging from dictionaries (Grefenstette, 1994:81), lexical databases like *WordNet* (Lin, 1998) to experimental data as obtained from humans (Landauer & Dumais, 1997). However, when looking at the literature, it seems that the data from the synonym portion of the *Test of English as a Foreign Language* (TOEFL) has gained some dominance for the evaluation of semantic relatedness. It has the advantage that it is easy to use, easy to interpret, and that it directly reflects human intuition. In comparison, data from dictionaries or lexical databases has been produced using sophisticated processes and therefore may be more distant from spontaneous human intuition.

The TOEFL data has been first used by Landauer & Dumais (1997) who obtained it from the *Educational Testing Service*. The TOEFL is an obligatory test for non-native speakers of English who would like to study at a university with English as the teaching language. The data used by Landauer & Dumais comprises 80 test items. Each item consists of a problem word in testing parlance and four alternative words, from which the test taker is asked to choose the one with the most similar meaning to the problem word. For example, given the test sentence "*Both boats and trains are used for transporting the materials*" and the four alternative words *planes*, *ships*, *canoes*, and *railroads*, the subject would be expected to choose the word *ships*, which is supposed to be the one most similar to *boats*.

Landauer & Dumais (1997) found that their algorithm for computing semantic similarities between words, which is based on an SVD-approach called *latent semantic analysis*, has a similar success rate when applied to the TOEFL synonym test as the human test takers. Whereas the algorithm got 64.4% of the questions right, the success rate of the human subjects was 64.5%. Other researchers were able to improve the performance to 69% (Rapp, 2002), 72% (Sahlgren, 2001), 74% (Turney, 2001) and 81.25% (Terra & Clarke, 2003). This gives the impression that the quality of the simulation is above human level.

However, it has often been overlooked that the 64.5% performance figure achieved by the test takers relates to non-native speakers of English, and that native speakers would perform significantly better. On the other hand, the simulation programs are usually not designed to make use of the context of the test word, so they neglect some information that may be useful for the human subjects.

In order to approach both issues, we have presented the TOEFL test words, together with the alternative words, but without the sentences, to five native and five non-native speakers of English, drawn from staff at Macquarie University (Sydney), and recorded their choices. Two of the native speakers got all 80 items correct, another two got 78 correct, and one got 75 correct. Table 1 gives an overview of the errors made by the native speakers. As expected, the performance of the non-native speakers was considerably worse. Their numbers of correct choices were 75, 70, 69, 67, and 66.

TEST WORDS	ALTERNATIVE WORDS
expeditiously	frequently actually (1) *rapidly* repeatedly
fashion	ration fathom craze (1) *manner*
figure	list *solve* divide express (1)
issues	training salaries *subjects* benefits (1)
levied	*imposed* believed requested correlated (1)
showy	*striking* prickly entertaining (3) incidental
wildly	distinctively mysteriously abruptly (1) *furiously*

Table 1. Errors of the native speakers. The correct choices are set in italics, and the number of subjects who selected a wrong answer follows the respective word in brackets.

On average, the performance of the native speakers was 97.75%, whereas the performance of the non-native speakers was 86.75%. Remember that the performance of the non-native speakers in the TOEFL test, although they had the context of each test word as an additional clue, was only 64.5%. The discrepancy of more than 20% between our non-native speakers and the TOEFL test takers can be explained by the fact that most of our subjects had spent many years in English speaking countries and thus had a language proficiency far above average. More importantly, our native speakers' results indicate that the performance of the above mentioned algorithms is clearly below human performance. So the impression from the Landauer & Dumais (1997) paper and from the information retrieval literature (Ruge, 1992) that human-like quality has already been achieved is obviously wrong unless one only looks at second language learners with a relatively poor proficiency.

3. Algorithm

Finding that there is a lot of room for improvement, we used the algorithm described in the seminal paper by Landauer & Dumais (1997) as a basis and modified all its details in a systematic way. The changes include using a larger and more balanced corpus, lemmatizing it, modifying the window type and size, as well as adapting the association formula and the dimensionality of the matrix. Since all parameters influence each other and cannot simply be optimized separately, the difficult part in doing so is to find the right balance. The resulting algorithm is briefly described below. Some more detail is given in another paper (Rapp, 2003).

Since our aim is to simulate human intuitions on word meanings by analyzing the statistical distribution of words in a large text collection, it is important that this collection represents a balanced sample of different varieties of language use. This requirement is best fulfilled by the British National Corpus (BNC). Note, however, that with 100 million words this corpus is much smaller than some newspaper corpora or, for example, the terabyte web corpus (53 billion words) used by Terra & Clarke (2003). Due to restrictions in vocabulary size imposed by the SVD, we (partially) lemmatized the BNC on the basis of a lexicon provided by Karp et al. (1992). That is, all word forms that (without considering context) could be unambiguously assigned to a lemma were replaced by the root form. Note that this process introduces some errors, as the lexicon is not error free and not complete. The same lemmatization procedure was also applied to the TOEFL test data. In another step, based on a list of approximately 200 items, we removed the function words from the BNC.

Using a window size of ±2, we then computed a co-occurrence matrix from the pre-processed corpus. By storing it as a sparse matrix, it was feasible to include all of the approximately 375000 lemmas occurring in the BNC.

Although semantic similarities can be successfully computed based on raw word co-occurrence counts, the results can be improved when the observed co-occurrence-frequencies are transformed by some function that reduces the effects of different word frequencies. As motivated in Rapp (2003), we use here a modified version of the entropy-based transformation functions described by Landauer & Dumais (1997):

$$A_{ij} = \log(1+f_{ij}) \cdot \left(-\sum_{k} p_{kj} \log(p_{kj}) \right) \quad \text{with} \quad p_{kj} = \frac{f_{kj}}{c_j}$$

Here, f_{ij} is the co-occurrence frequency of words i and j and c_j is the corpus frequency of word j. Indices i, j, and k all have a range between one and the vocabulary size n. The sum term in the formula is entropy. As usual with entropy, it is assumed that $0 \log(0) = 0$.

Let us now look at how the formula works. The important part is taking the logarithm of f_{ij} thus dampening the effects of large differences in frequency. Adding 1 to f_{ij} provides some smoothing and prevents the logarithm from becoming infinite if f_{ij} is zero. Some improvement can be achieved by multiplying this by the entropy of a word. This has the effect that the weights of rare words with only few (and often incidental) co-occurrences are reduced.

Following the findings by Schütze (1997) and Landauer & Dumais (1997) we reduced the number of dimensions of the resulting association matrix by applying the SVD. This way some smoothing and generalization effect can be achieved that has been shown to improve the results of subsequent similarity computations.

However, since the SVD is computationally rather demanding, before applying it we first removed all words with a corpus frequency below 20 from the matrix. This reduced its size from 374244 × 374244 to 56096 × 56096. By using a version of Mike Berry's SVDPACK software that had been modified and kindly provided by Hinrich Schütze, we transformed this smaller matrix to a matrix of 56096 lines and 300 columns. The resulting dimensionality-reduced matrix has not only the advantage that all subsequent similarity computations are much faster, but also that the results tend to agree better with human intuitions on word similarity.

To determine the words most similar to a given word, its vector (line in the matrix) is compared to the vectors of all other words in the matrix using one of the standard vector similarity measures. Those words that obtain the best scores are considered to be most similar. Among the many possible similarity measures found in the literature we chose the cosine coefficient which works very well in conjunction with SVD-processed matrices (Rapp, 2003). It computes the cosine of the angle between two vectors.

4. Results and evaluation

To give a first impression, table 2 shows the top most similar words to a few examples as computed using SVD, the cosine-coefficient, and a vocabulary of 56096 words. Although these results look plausible, a quantitative evaluation is always desirable. For this reason we used our system for solving the TOEFL synonym test and compared the results to the correct answers as provided by the Educational Testing Service. Remember that the subjects had to choose the word most similar to a given stimulus word from a list of four alternatives. In the simulation, we assumed that the system made the right decision if the correct answer was ranked highest among the four alternatives. This was the case for 74 of the 80 test items which gives us an accuracy of 92.5%. For comparison, recall that the performance of our human subjects had been 97.75% for the native speakers and 86.75% for our highly proficient non-native speakers. This means our program's performance is in between the two levels with about equal margins towards both sides.

Let us now have a look at the six errors that the program made (table 3). In some cases the program's "reasoning" seems fairly obvious. For *fanciful*, it chose *logical* because this word has an antonymie character. For *figure*, which can be a noun or a verb, *list* was chosen instead of *solve* because in the BNC the use as a noun prevails. In the case of *halfheartedly* the program's choice becomes guesswork as this word as well as the correct answer *apathetically* occur only a few times in the BNC.

enormously	greatly (0.52) immensely (0.51) tremendously (0.48) considerably (0.48) substantially (0.44) vastly (0.38) hugely (0.38) dramatically (0.35)
flaw	shortcomings(0.43) defect (0.42) deficiencies (0.41) weakness (0.41) fault (0.36) drawback (0.36) anomaly (0.34) inconsistency (0.34)
issue	question (0.51) matter (0.47) debate (0.38) concern (0.38) problem (0.37) topic (0.34) consideration (0.31) raise (0.30) dilemma (0.29)
build	building (0.55) construct (0.48) erect (0.39) design (0.37) create (0.37) develop (0.36) construction (0.34) rebuild (0.34) exist (0.29)
discrepancy	disparity (0.44) anomaly (0.43) inconsistency (0.43) inaccuracy (0.40) difference (0.36) shortcomings (0.35) variance (0.34) imbalance (0.34)
essentially	primarily (0.50) largely (0.49) purely (0.48) basically (0.48) mainly (0.46) mostly (0.39) fundamentally (0.39) principally (0.39) solely (0.36)

Table 2. Semantic similarities as computed. The lists are ranked according to the cosine coefficient.

TEST WORDS	ALTERNATIVE WORDS
fanciful	familiar *imaginative* apparent LOGICAL
figure	LIST *solve* divide express
halfheartedly	CUSTOMARILY bipartisanly *apathetically* unconventionally
provisions	*stipulations* interrelations jurisdictions INTERPRETATIONS
roots	*origins* rituals CURE function
temperate	COLD *mild* short windy

Table 3. Errors made by the program. (Expected answers are set in italics, simulation results in small capitals.)

5. Comparison with other systems

In section 2, the performances of some other systems that also had been evaluated on the TOEFL synonym test have been given. The best performance we are aware of was reported by Terra & Clarke (2003) which is 81.25%. It was obtained using a terabyte web corpus (53 billion words) and pointwise mutual information as the similarity measure. Although our corpus is several orders of magnitude smaller, by using SVD with a performance of 92.5% we were able to significantly improve on this result, which is an indication that the generalization effect claimed for SVD actually works in practice. This finding is also supported by our previous performance of only 69% achieved on the BNC without SVD (Rapp, 2002).

Unfortunately, Dekang Lin's well known dependency-based thesaurus of similar words was not evaluated on the TOEFL data, but instead a sophisticated comparison with WordNet was conducted (Lin, 1998). As this thesaurus is available on the web (http://www.cs.ualberta.ca/~lindek/demos/depsim.htm), its high quality is easily verifiable. According to Lin (1998), Grefenstette (1994), and Ruge (1992), word similarities should not be computed on the basis of the co-occurrences of all words as found in raw text, but instead by using a shallow or a full parser only dependency relations should be extracted. The view here is that the window-based methods may work to some extent, but that many of the word co-occurrences in a text window are incidental and only add noise to the significant word pairs. Also, parsing can help to resolve those types of semantic ambiguities that have an effect on syntax. A good example is the word *sound* which can be an adjective, a verb, or a noun and in each case has a different meaning.

To find out to what extent parsing improves results, we evaluated Lin's dependency-based thesaurus (as found on the internet in February 2004) on the TOEFL data. For each of the 80 test words, in the ranked lists of the thesaurus we looked up the positions of the four alternative words. As for three test items either no data was available (test item *hind*) or none of the alternative words appeared in the lists (test items *showy* and *uniform*), we reduced our test set to the 77 remaining items. In 7 of the 77 cases at least one of the incorrect alternatives was ranked ahead of the correct answer, i.e. the prediction of the system was wrong. This gives us an accuracy of 90.9%. Table 4 shows the predictions that Lin's thesaurus got wrong.

Note that in those cases where an ambiguous word can belong to more than one part of speech, in analogy to WordNet Lin's thesaurus offers separate lists for each possibility. We decided to count an item as being correctly answered if any of the lists made the correct prediction, even if other lists were incorrect. This gives Lin's thesaurus some advantage, as for words that are ambiguous concerning their part of speech it has more than one chance to get it right. On the other hand, Lin's algorithm had been optimized by comparing the results to WordNet and not to the TOEFL task, which in our setting can be seen as a disadvantage. Also, in a few cases it is not clear whether a wrong prediction has been caused by missing words in the vocabulary. We tried to minimize this effect by taking into account inflected or closely related forms of a word. For example, given the word *solitary*, we counted *lone* as being correct, although the expected answer *alone* was missing in the list.

TEST WORDS	ALTERNATIVE WORDS
furnish	*supply* impress PROTECT advise
temperate	cold *mild* short WINDY
situated	rotating ISOLATED emptying *positioned*
halfhearted(ly)	customarily BIPARTISAN(LY) *apathetical(ly)* unconventional(ly)
hailed	judged *acclaimed* REMEMBERED addressed
tranquillity	*peacefulness* harshness weariness HAPPINESS
feasible	permitted *possible* EQUITABLE evident

Table 4. Wrong choices made by Lin's thesaurus.

Putting these shortcomings aside, our comparison indicates that Lin's method does not perform better on the TOEFL task than our purely statistical SVD-based approach. The use of syntax therefore seems not essential for this purpose. However, especially for certain ambiguous words, Lin's similarity rankings look considerably less noisy than ours. Neglecting differences in corpus size and term segmentation, in our view it is not clear whether the main reason for this is that by parsing the corpus he explicitly identifies the dependency relations, or if it is the disambiguation effect that the parser provides in that it distinguishes between those senses of a word form that belong to different parts of speech.

Should the latter be true, then if no parser is available for a language, the same disambiguation effect could quite as well be achieved using a part-of-speech tagger. And in case even a part-of-speech tagger is unavailable, then simply filtering the output lists of our program according to part of speech would bring an improvement, as this ensures that only the desired paradigmatic relations remain.

However, for us the surprising thing about this research is not that the use of sophisticated linguistic tools can possibly bring an improvement, but that a fully unsupervised approach which only relies on algebra seems to come fairly close in performance. Neglecting the preprocessing step of partial lemmatization, which essentially served the purpose of keeping our co-occurrence matrix small enough for SVD processing, no linguistic resources, neither a lexicon nor syntactic rules, are required. The algorithm considers any string of characters that is delimited by blanks or punctuation marks as a word, applies the SVD to an association matrix derived from the co-occurrences of the words in a corpus, and finally comes up with lists of similar words that highly agree with human intuitions.

6. Summary and future work

We have presented a statistical method for the automatic computation of related words from a corpus which has been evaluated on the TOEFL synonym test. Its performance on this task favorably compares to other purely statistical approaches and suggests that sophisticated and language dependent syntactic processing is not essential.

The automatically generated English thesaurus of similar word comprising 56096 entries is freely available from the web at http://www.fask.uni-mainz.de/user/rapp/. A similar thesaurus for German is also available. Although, unlike other thesauri, at the current stage it does not distinguish between different kinds of relationships between words, it has one advantage over manually created thesauri: Given a certain word, it not only lists a few related words, but instead ranks all words of a large vocabulary according to their similarity to the given word. Since even at the higher ranks the distinctions obtained seem meaningful, this is an important feature that is indispensable for certain kinds of machine processing, e.g. for word sense disambiguation and induction.

Future work that we envisage includes applying our method to corpora from other languages, adding multiword units to the vocabulary, and to find solutions to the problem of word ambiguity that has not been dealt with in this work. If the scope is to be expanded from finding related words to other tasks that require a certain amount of syntactical processing, as for example necessary when trying to identify the different kinds of relationships between words, then a promising direction of research could be not to consider parsing and SVD as competitors, but to put the SVD on top of a syntax-based approach.

7. References

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Dordrecht: Kluwer.

Karp, D.; Schabes, Y.; Zaidel, M.; Egedi, D. (1992). A freely available wide coverage morphological analyzer for English. In: *Proceedings of 14th COLING,* Nantes, 950–955.

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

Lin, D. (1998). Automatic retrieval and clustering of similar words. In: *Proceedings of COLING-ACL 1998,* Montreal, Vol. 2, 768–773.

Rapp, R. (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. *Proc. of 19th COLING,* Taipei, ROC, Vol. 2, 821–827.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In: *Proceedings of the Ninth Machine Translation Summit,* New Orleans, 315–322.

Ruge, G. (1992). Experiments on linguistically based term associations. *Information Processing and Management* 28(3), 317–332.

Sahlgren, M. (2001). Vector-based semantic analysis: representing word meanings based on random labels. In: A. Lenci, S. Montemagni, V. Pirrelli (eds.): *Proceedings of the ESSLLI Workshop on the Acquisition and Representation of Word Meaning,* Helsinki.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Stanford: CSLI Publications.

Terra, E., Clarke, C.L.A. (2003). Frequency estimates for statistical word similarity measures. *Proceedings of HLT/NAACL*, Edmonton, Alberta, May 2003.

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

Acknowledgements

Part of this research was conducted during my stay at the Centre for Language Technology of Macquarie University, Sydney. I would like to thank Robert Dale and his research group, Manfred Wettler, Raz Tamir, and Gerda Ruge for their help and advice, Hinrich Schütze and Mike Berry for the SVD software, and the DFG for financially supporting this work.