The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches

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

Abstract

It is shown that basic language processes such as the production of free word associations and the generation of synonyms can be simulated using statistical models that analyze the distribution of words in large text corpora. According to the law of association by contiguity, the acquisition of word associations can be explained by Hebbian learning. The free word associations as produced by subjects on presentation of single stimulus words can thus be predicted by applying first-order statistics to the frequencies of word co-occurrences as observed in texts. The generation of synonyms can also be conducted on co-occurrence data but requires second-order statistics. The reason is that synonyms rarely occur together but appear in similar lexical neighborhoods. Both approaches are systematically compared and are validated on empirical data. It turns out that for both tasks the performance of the statistical system is comparable to the performance of human subjects.

1 Introduction

According to Ferdinand de Saussure (1916), there are two fundamental types of relations between words that he believes correspond to basic operations of our brain: *syntagmatic* and *paradigmatic* associations. There is a syntagmatic relation between two words if they co-occur in spoken or written language more frequently than expected from chance and if they have different grammatical roles in the sentences in which they occur. Typical examples are the word pairs *coffee – drink*, *sun – hot*, or *teacher – school*. The relation between two words is paradigmatic if the two words can substitute for one another in a sentence without affecting the grammaticality or acceptability of the sentence. Typical examples are synonyms or antonyms like *quick – fast*, or *eat – drink*. Normally, words with a paradigmatic relation are the same part of speech, whereas words with a syntagmatic relation can but need not be the same part of speech.

In this paper we want to show that the two types of relations as defined by de Saussure are reflected in the statistical distribution of words in large corpora. We present algorithms that automatically retrieve words with either the syntagmatic or the paradigmatic type of relationship from corpora and perform a quantitative evaluation of our results.

2 Paradigmatic Associations

Paradigmatic associations are words with high semantic similarity. According to Ruge (1992), 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 *red* and *blue* can be derived from the fact that they both frequently co-occur with words like *color, flower, dress, car, dark, bright, beautiful,* and so forth. If for each word in a corpus a co-occurrence vector is determined whose entries are the co-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 using one of the standard similarity measures, for example, the cosine coefficient. Those words that obtain the best values are considered to be most similar. Practical implementations of algorithms based on this principle have led to excellent results as documented in papers by Ruge (1992), Grefenstette (1994), Agarwal (1995), Landauer & Dumais (1997), Schütze (1997), and Lin (1998).

2.1 Human Data

In this section we relate the results of our version of such an algorithm to similarity estimates obtained by human subjects. Fortunately, we did not need to conduct our own experiment to obtain the human's similarity estimates. Instead, such data was kindly provided by Thomas K. Landauer, who had taken it from the synonym portion of the *Test of English as a Foreign Language* (TOEFL). Originally, the data came, along with normative data, from the Educational Testing Service (Landauer & Dumais 1997). The TOEFL is an obligatory test for foreign students who would like to study at an American or English university.

The data 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 that with the most similar meaning to the problem word. For example, given the test sentence "*Both <u>boats</u> 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 the one most similar to *boats*.

2.2 Corpus

As mentioned above, our method of simulating this kind of behavior is based on regularities in the statistical distribution of words in a corpus. We chose to use the British National Corpus (BNC), a 100-million-word corpus of written and spoken language that was compiled with the intention of providing a representative sample of British English.

Since this corpus is rather large, to save disk space and processing time we decided to remove all function words from the text. This was done on the basis of a list of approximately 200 English function words. We also decided to lemmatize the corpus as well as the test data. This not only reduces the sparse-data problem but also significantly reduces the size of the co-occurrence matrix to be computed. More details on these two steps of corpus preprocessing can be found in Rapp (1999).

2.3 Co-occurrence Counting

For counting word co-occurrences, as in most other studies a fixed window size is chosen and it is determined how often each pair of words occurs within a text window of this size. Choosing a window size usually means a trade-off between two parameters: specificity versus the sparse-data problem. The smaller the window, the stronger the associative relation between the words inside the window, but the more severe the sparse data problem (see figure 1 in section 3.2). In our case, with ±1 word, the window size looks rather small. However, this can be justified since we have reduced the effects of the sparse-data problem by using a large corpus and by lemmatizing the corpus. It also should be noted that a window size of ±1 applied after elimination of the function words is comparable to a window size of ±2 without elimination of the function words (assuming that roughly every second word is a function word).

Based on the window size of ±1, we computed a co-occurrence matrix of about a million words in the lemmatized BNC. Although the resulting matrix is extremely large, this was feasible since we used a sparse format that does not store zero entries.

2.4 Computation of Word Similarities

To determine the words most similar to a given word, the co-occurrence vector of this word is compared to all other vectors in the matrix and the words are ranked according to the similarity values obtained. It is expected that the most similar words are ranked first in the sorted list.

For vector comparison, different similarity measures can be considered. Salton & McGill (1983) proposed a number of measures, such as the cosine coefficient, the Jaccard coefficient, and the Dice coefficient. For the computation of related terms and synonyms, Ruge (1995) and Landauer & Dumais (1997) used the cosine measure, whereas Grefenstette (1994, p. 48) used a weighted Jaccard measure. We propose here the city-block metric, which computes the similarity between two vectors *X* and *Y* as the sum of the absolute differences of corresponding vector positions:

$$s = \sum_{i=1}^{n} \left| X_i - Y_i \right|$$

In a number of experiments we compared it to other similarity measures, such as the cosine measure, the Jaccard measure (standard and binary), the Euclidean distance, and the scalar product, and found that the city-block metric yielded good results (see Rapp, 1999).

2.5 Results

Table 1 shows the top five paradigmatic associations to six stimulus words. As can be seen from the table, nearly all words listed are of the same part of speech as the stimulus word. Of course, our definition of the term *paradigmatic association* as given in the introduction implies this. However, the simulation system never obtained any information on part of speech, and so it is nevertheless surprising that – besides computing term similarities – it implicitly seems to be able to cluster parts of speech. This observation is consistent with other studies (e.g., Ruge, 1995).

blue	cold	fruit	green	tobacco	whiskey
red	hot	food	red	cigarette	whisky
green	warm	flower	blue	alcohol	brandy
grey	dry	fish	white	coal	champagne
yellow	drink	meat	yellow	import	lemonade
white	cool	vegetable	grey	textile	vodka

Table 1: Computed paradigmatic associations.

A qualitative inspection of the word lists generated by the system shows that the results are quite satisfactory. Paradigmatic associations like *blue* → *red*, *cold* → *hot*, and *tobacco* → *cigarette* are intuitively plausible. However, a quantitative evaluation would be preferable, of course, and for this reason we did a comparison with the results of the human subjects in the TOEFL test. Remember that the human 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 had chosen the correct alternative if the correct word was ranked highest among the four alternatives. This was the case for 55 of the 80 test items, which gives us an accuracy of 69%. This accuracy may seem low, but it should be taken into account that the TOEFL tests the language abilities of prospective university students and therefore is rather difficult. Actually, the performance of the average human test taker was worse than the performance of the system. The human subjects were only able to solve 51.6 of the test items correctly, which gives an accuracy of 64.5%. Please note that in the TOEFL, average performance (over several types of tests, with the synonym test being just one of them) admits students to most universities. On the other hand, by definition, the test takers did not have a native command of English, so the performance of native speakers would be expected to be significantly better. Another consideration is the fact that our simulation program was not designed to make use of the context of the test word, so it neglected some information that may have been useful for the human subjects.

Nevertheless, the results look encouraging. Given that our method is rather simple, let us now compare our results to the results obtained with more sophisticated methods. One of the methods reported in the literature is singular value decomposition (SVD); another is shallow parsing. SVD, as described by Schütze (1997) and Landauer & Dumais (1997), is a method similar to factor analysis or multi-dimensional scaling that allows a significant reduction of the dimensionality of a matrix with minimum information loss. Landauer & Dumais (1997) claim that by optimizing the dimensionality of the target matrix the performance of their word similarity predictions was significantly improved.

However, on the TOEFL task mentioned above, after empirically determining the optimal dimensionality of their matrix, they report an accuracy of 64.4%. This is somewhat worse than our result of 69%, which was achieved without SVD and without optimizing any parameters. It must be emphasized, however, that the validity of this comparison is questionable, as many parameters of the two models are different, making it unclear which ones are responsible for the difference. For example, Landauer and Dumais used a smaller corpus (4.7 million words), a larger window size (151 words on average), and a different similarity measure (cosine measure). We nevertheless tend to interpret the results of our comparison as evidence for the view that SVD is just another method for smoothing that has its greatest benefits for sparse data. However, we do not deny the technical value of the method. The one-time effort of the dimensionality reduction may be well spent in a practical system because all subsequent vector comparisons will be speeded up considerably with shorter vectors.

Let us now compare our results to those obtained using shallow parsing, as previously done by Grefenstette (1993). The view here is that the window-based method may work to some extent, but that many of the word co-occurrences in a window are just incidental and add noise to the significant word pairs. A simple method to reduce this problem could be to introduce a threshold for the minimum number of co-occurrences; a more sophisticated method is the use of a (shallow) parser. Ruge (1992), who was the first to introduce this method, claims that only head-modifier relations, as known from dependency grammar, should be considered. For example, if we consider the sentence "Peter drives the blue car", then we should not count the co-occurrence of *Peter* and *blue,* because *blue* is neither head nor modifier of *Peter*. Ruge developed a shallow parser that is able to determine the head-modifier relations in unrestricted English text with a recall of 85% and a precision of 86% (Ruge, 1995). Using this parser she extracted all head-modifier relations from the 100 million words of the British National Corpus. Thus, the resulting co-occurrence matrix only contained the counts of the head-modifier relations. The word similarities were computed from this matrix by using the cosine similarity measure. Using this method, Ruge achieved an accuracy of about 69% in the TOEFL synonym task, which is equivalent to our results.

Again, we need to emphasize that parameters other than the basic methodology could have influenced the result, so we need to be cautious with an interpretation. However, to us it seems that the view that some of the co-occurrences in corpora should be considered as noise is wrong, or else if there is some noise it obviously cancels out over large corpora. It would be interesting to know how a system performed that used all co-occurrences except the head-modifier relations. We tend to assume that such a system would perform worse, so the parser selected the good candidates. However, the experiment has not been done, so we cannot be sure.

Although the shallow parsing could not improve the results in this case, we nevertheless should point out its virtues: It improves efficiency since it leads to sparser matrices. It also seems to be able to separate the relevant from the irrelevant co-occurrences. Third, it may be useful for determining the type of relationship between words (e.g., synonymy, antonymy, meronymy, hyponymy, etc., see Berland & Charniak, 1999). Although this is not within the scope of this paper, it is very relevant for related tasks, for example, the automatic generation of thesauri.

3 Syntagmatic Associations

Syntagmatic associations are words that frequently occur together. Therefore, an obvious approach to extract them from corpora is to look for word pairs whose co-occurrence is significantly larger than chance. To test for significance, the standard chi-square test can be used. However, Dunning (1993) pointed out that for the purpose of corpus statistics, where the sparseness of data is an important issue, it is better to use the log-likelihood ratio. It would then be assumed that the strongest syntagmatic association to a word would be that other word that gets the highest log-likelihood score.

Please note that this method is computationally far more efficient than the computation of paradigmatic associations. For the computation of the syntagmatic associations to a stimulus word only the vector of this single word has to be considered, whereas for the computation of paradigmatic associations the vector of the stimulus word has to be compared to the vectors of all other words in the vocabulary. The computation of syntagmatic associations is said to be of first-order type, whereas the computation of paradigmatic associations is of second-order type. Algorithms for the computation of first-order associations have been used in lexicography for the extraction of collocations (Smadja, 1993) and in cognitive psychology for the simulation of associative learning (Wettler & Rapp, 1993).

3.1 Association Norms

As we did with the paradigmatic associations, we would like to compare the results of our simulation to human performance. However, it is difficult to say what kind of experiment should be conducted to obtain human data. As with the paradigmatic associations, we decided not to conduct our own experiment but to use the Edinburgh Associative Thesaurus (EAT), a large collection of association norms, as compiled by Kiss et al. (1973). Kiss presented lists of stimulus words to human subjects and asked them to write after each word the first word that the stimulus word made them think of. Table 2 gives some examples of the associations the subjects came up with.

As can be seen from the table, not all of the associations given by the subjects seem to be of syntagmatic type. For example, the word pairs *blue* –

black or *cold – hot* are clearly of paradigmatic type. This observation is of importance and will be discussed later.

blue	cold	fruit	green	tobacco	whiskey
sky	hot	apple	grass	smoke	drink
black	ice	juice	blue	cigarette	gin
green	warm	orange	red	pipe	bottle
red	water	salad	yellow	poach	soda
white	freeze	machine	field	road	Scotch

Table 2: Some sample associations from the EAT.

3.2 Computation

For the computation of the syntagmatic associations we used the same corpus as before, namely the British National Corpus. In a preliminary experiment we tested if there is a correlation between the occurrence of a stimulus word in the corpus and the occurrence of the most frequent associative response as given by the subjects. For this purpose, we selected 100 stimulus/response pairs and plotted a bar chart from the co-occurrence data (see figure 1). In the bar chart, the x-axis corresponds to the distance of the response word from the stimulus word (measured as the number of words separating them), and the y-axis corresponds to the occurrence frequency of the response word in a particular distance from the stimulus word. Please note that for the purpose of plotting this bar chart, function words have been taken into account.

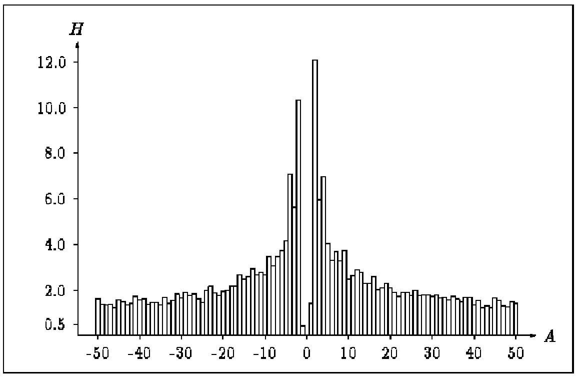

Figure 1: Occurrence frequency *H* of a response word in a particular distance *A* from the corresponding stimulus word (averaged over 100 stimulus/response pairs).

As can be seen from the figure, the closer we get to the stimulus word, the more likely it is that we find an occurrence of its strongest associative response. Exceptions are the positions directly neighboring the stimulus word. Here it is rather unlikely to find the response word. This observation can be explained by the fact that content words are most often separated by function words, so that the neighboring positions are occupied by function words.

Now that it has been shown that there is some relationship between human word associations and word co-occurrences, let us briefly introduce our algorithm for extracting word associations from texts. Based on a window size of ±20 words, we first compute the co-occurrence vector for a given stimulus word, thereby eliminating all words with a corpus frequency of less than 101. We then apply the log-likelihood test to this vector. According to Lawson & Belica[1] the log-likelihood ratio can be computed as follows: Given the word *W*, for each co-occurring word *S*, its window frequency *A*, its residual frequency *C* in the reference corpus, the residual window size *B* and the residual corpus size *D* are stored in a 2 by 2 contingency table.

	S	¬*S*	Total
W	*A*	*B*	*A*+*B*
¬*W*	*C*	*D*	*C*+*D*
Total	*A*+*C*	*B*+*D*	*N*

Then the log-likelihood statistics are calculated:

$$\begin{aligned} G = {} & 2(A \log A + B \log B + C \log C + D \log D \\ & + N \log N - (A+B) \log(A+B) \\ & - (A+C) \log(A+C) - (B+D) \\ & \log(B+D) - (C+D) \log(C+D)) \end{aligned}$$

Finally, the vocabulary is ranked according to descending values of *G* as computed for each word. The word with the highest value is considered to be the primary associative response.

3.3 Results

In table 3 a few sample association lists as predicted by our system are listed. They can be compared to the human associative responses given in table 2.

The valuation of the predictions has to take into account that association norms are conglomerates of the answers of different subjects that differ considerably from each other. A satisfactory prediction would be proven if the difference between the pre-

[1] Handout at GLDV Meeting, Frankfurt/Main 1999.

dicted and the observed responses were about equal to the difference between an average subject and the rest of the subjects. This is actually the case. For 27 out of the 100 stimulus words the predicted response is equal to the observed primary response. This compares to an average of 28 primary responses given by a subject in the EAT. Other evaluation measures lead to similar good results (Wettler & Rapp, 1993; Rapp, 1996).

blue	cold	fruit	green	tobacco	whiskey
red	hot	vegetable	red	advertising	drink
eyes	water	juice	blue	smoke	Jesse
sky	warm	fresh	yellow	ban	bottle
white	weather	tree	leaves	cigarette	Irish
green	winter	salad	colour	alcohol	pour

Table 3: Results with the co-occurrence-based approach.

We conclude from this that our method seems to be well suited to predict the free word associations as produced by humans. And as human associations are not only of syntagmatic but also of paradigmatic type, so does the co-occurrence-based method predict both types of associations rather well. In the ranked lists produced by the system we find a mixture of both types of associations. However, for a given association there is no indication whether it is of syntagmatic or paradigmatic type.

We suggest a simple method to distinguish the paradigmatic from the syntagmatic associations. Remember that the 2nd-order approach described in the previous section produced paradigmatic associations only. So if we simply remove the words produced by the 2nd-order approach from the word lists obtained by the 1st-order approach, then this should give us solely syntagmatic associations.

4 Comparison between Syntagmatic and Paradigmatic Associations

Table 4 compares the top five associations to a few stimulus words as produced by the 1st-order and the 2nd-order approach. In the list, we have printed in bold those 1st-order associations that are not among the top five in the second-order lists. Further inspections of these words shows that they are all syntagmatic associations. So the method proposed seems to work in principle. However, we have not yet conducted a systematic quantitative evaluation. Conducting a systematic evaluation is not trivial, since the definitions of the terms *syntagmatic* and *paradigmatic* as given in the introduction may not be precise enough. Also, for a high recall, the word lists considered should be much longer than the top five. However, the further down we go in the ranked lists, the less typical are the associations. So it is not clear where to automatically set a threshold. We did not further elaborate on this because for our practical work this issue was of lesser importance.

Although both algorithms are based on word co-occurrences, our impression is that their strengths and weaknesses are rather different. So we see a good chance of obtaining an improved generator for associations by combining the two methods.

stimulus	1st-order	2nd-order
blue	*red*	*red*
	eyes	*green*
	sky	grey
	white	yellow
	green	*white*
cold	*hot*	*hot*
	water	*warm*
	warm	dry
	weather	drink
	winter	cool
fruit	*vegetable*	food
	juice	flower
	fresh	fish
	tree	meat
	salad	*vegetable*
green	*red*	*red*
	blue	*blue*
	yellow	white
	leaves	*yellow*
	colour	grey
tobacco	**advertising**	*cigarette*
	smoke	*alcohol*
	ban	coal
	cigarette	import
	alcohol	textile
whiskey	**drink**	whisky
	Jesse	brandy
	bottle	champagne
	Irish	lemonade
	pour	vodka

Table 4: Comparison between 1st-order and 2nd-order associations.

5 Discussion and Conclusion

We have described algorithms for the computation of 1st-order and 2nd-order associations. The results obtained have been compared with the answers of human subjects in the free association task and in the TOEFL synonym test. It could be shown that the performance of our system is comparable to the performance of the subjects for both tasks.

We observed that there seems to be some relationship between the type of computation performed (1st-order versus 2nd-order) and the terms *syntagmatic* and *paradigmatic* as coined by de Saussure. Whereas the results of the 2nd-order computation are of paradigmatic type exclusively, those of the 1st-order computation are a mixture of both syntagmatic and paradigmatic associations. Removing the 2nd-order associations from the 1st-order associations leads to solely syntagmatic associations.

We believe that the observed relation between our statistical models and the intuitions of de Saussure are not incidental, and that the striking similarity of the simulation results with the human associations also has a deeper reason. Our explanation for this is that human associative behavior is governed by the law of association by contiguity, which is well known from psychology (Wettler, Rapp & Ferber, 1993). In essence, this means that in the process of learning or generating associations the human mind seems to conduct operations that are equivalent to co-occurrence counting, to performing significance tests, or to computing vector similarities (see also Landauer & Dumais, 1997). However, further work is required to find out to what extent other language-related tasks can also be explained statistically.

Acknowledgements

This research was supported by the DFG. I would like to thank Manfred Wettler and Gerda Ruge for many inspiring discussions.

References

Agarwal, R. (1995). *Semantic Feature Extraction from Technical Texts with Limited Human Intervention.* Dissertation, Mississippi State University.

Berland, M., Charniak, E. (1999). Finding Parts in Very Large Corpora. In: *Proceedings of ACL 1999,* College Park. 57–64.

de Saussure, F. (1916/1996). *Cours de linguistique générale.* Paris: Payot.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.

Grefenstette, G. (1993). Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In: *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, Ohio.

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

Kiss, G.R., Armstrong, C., Milroy, R., Piper, J. (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley and N. Hamilton-Smith (eds.): *The Computer and Literary Studies*. Edinburgh: University Press.

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. (1996). *Die Berechnung von Assoziationen.* Hildesheim: Olms.

Rapp, R. (1999). Automatic identification of word translation from unrelated English and German corpora. In: *Proceedings of ACL 1999,* College Park. 519–526.

Ruge, G. (1992). Experiments on Linguistically Based Term Associations. *Information Processing & Management* 28(3), 317–332.

Ruge, G. (1995). *Wortbedeutung und Termassoziation.* Hildesheim: Olms.

Salton, G.; McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

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

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1), 143–177.

Wettler, M.; Rapp, R. (1993). Computation of word associations based on the co-occurrences of words in large corpora. In: *Proceedings of the 1st Workshop on Very Large Corpora:* Columbus, Ohio, 84–93.

Wettler, M., Rapp, R., Ferber, R. (1993). Freie Assoziationen und Kontiguitäten von Wörtern in Texten. *Zeitschrift für Psychologie*, 201, 99–108.