FI MU

Faculty of Informatics
Masaryk University

Finding Semantically Related Words in Large Corpora

by

Pavel Smrž
Pavel Rychlý

FI MU Report Series FIMU-RS-2001-02

June 2001

Finding Semantically Related Words in Large Corpora

Pavel Smrž and Pavel Rychlý

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {smrz,pary}@fi.muni.cz

Abstract. The paper deals with the linguistic problem of fully automatic grouping of semantically related words. We discuss the measures of semantic relatedness of basic word forms and describe the treatment of collocations. Next we present the procedure of hierarchical clustering of a very large number of semantically related words and give examples of the resulting partitioning of data in the form of dendrogram. Finally we show a form of the output presentation that facilitates the inspection of the resulting word clusters.

1 Introduction

The task of automatic finding of semantically related words belongs to the class of automatic lexical acquisition problems that attracted attention of many researchers in the area of natural language processing in last decades [1–3]. The term "semantical relatedness" denotes large group of language phenomena ranging from specific phenomena like synonyms, antonyms, hyperonyms, hyponyms, meronyms, etc. to more general ones, e.g. sets of words used in a particular scientific field or domain. In this paper, the task is understood in the wide sense.

The aim of finding groups of semantically related words is linguistically motivated by the assumption that semantically related words behave similarly. Information about semantic relatedness of a particular group of words is valuable for humans – consider for example foreign language learning and dictionaries arranged according to the topics. However, the strongest demand comes from the field of automatic natural language processing as it is one of the key issues in the solution of many problems in the field, namely the problem of selectional preferences or restrictions on particular type of verb arguments, in word sense disambiguation tasks, in machine translation, document classification, information retrieval and others.

The question remains how to cluster words according to the semantic domains or topics. The answer is motivated by the understanding of the task we have adopted above. Using the definition from [4] words are semantically similar (related) if they refer to entities in the world that are likely to co-occur. The simplest solution can therefore be based on the assumption that the words denoting such entities are also likely to co-occur in documents and it suffices to identify these words.

The first encountered problem when applying this strategy is the frequent coincidence of genuine semantic relatedness with the collocations in the result. The topic of collocation filtering is discussed in the following section.

The other problem concerns the fact that semantically related words do not need to co-occur in the same document. For example, Manning and Schütze [4] present terms *cosmonaut* and *astronaut* as the example of words that are not similar in the document space (they do not occur in the same documents) but are similar in the word space (they occur with the same words).[1]

The automatic lexical acquisition has been thoroughly studied in the field of corpus linguistics (see e.g. Boguraev and Pustejovsky [5]). The problem of semantic relatedness has been approached from the word co-occurrence statistics as well as from syntactic point of view [6]. There are also works on automatic enhancement of semantic hierarchies that can be viewed as a contribution to the semantic relatedness problem solution. The standard reference of the retrieving collocations from text is the work by Smajda [7].

The work most similar to ours is discussed in [4]. Manning and Schütze use logarithmic weighting function $f(x) = 1 + log(x)$ for non-zero co-occurrence counts, 25-word context window and cosine measure of semantic similarity. Unlike to our experiments, they compiled only some 1,000 most frequent words for so-called focus words and searched for about 20,000 most frequent words to form the word-by-word matrix. Moreover, the experiment described in [4] was aimed at automatic finding of the words that were most similar to the selected focus words. On the other hand, we present the method for automatic clustering of huge amount of frequently occurring words according to their semantic relatedness.

2 Prerequisites

2.1 How To Measure Semantic Relatedness

In the previous section we have defined the object of our interest – semantically related words – as words (not embodied in collocations) that are likely to co-occur within similar context. This section discusses how to characterize the fact that the words co-occur "frequently".

Several different methods have been applied to describe the notion of frequency. Statistical tests that define the probability of events co-occurrence are the most widely used. The t-test (or score), closely related z-score, or Pearson's

[1] It seems that the mentioned example does not work today in the time of world co-operation in space missions, and especially in the time of space partnership between Russians and Americans, as can be demonstrated by the corpus sentence: *A part of this project will be joined missions of Russian cosmonauts and American astronauts.* Notwithstanding this fact, we retain this example for its illustrativeness.

χ^2 (chi-square) test [8] belong to this category. The well-known likelihood ratio, that moreover takes advantage of clear interpretation, can also serve as a good characterization for these purposes, especially in the case of sparse data. Besides these statistically motivated measures we can apply the instrument of information theory, namely MI – (pointwise) mutual information – a measure that represents the amount of information provided by the occurrence of one entity about the occurrence of the other entity [4].

It has been shown many times that none of this measures works very well for low-frequency entities. For this reason, we have to exclude the low-frequency events from our observation. We have defined pragmatically motivated thresholds for minimal numbers of occurrences of examined events. As we have dealt with a huge amount of corpus data (approximately 100 millions of words), the restriction means no considerable limitation. Moreover, the concentration on the high-frequent events effaces the differences among various measures and decreases the dependency of the output quality on the choice of a particular measure. The mutual information measure used in our experiments gives similar results when compared with other methods and the problems with MI referred elsewhere [9] does not emerge.

2.2 Context Definition

The other important point in the definition of our goal is what we will understand by the notion "co-occurrence in the context". Context is straightforwardly defined in the area of information retrieval – it is given by means of documents. This approach is applicable also in the field of corpus linguistics as the majority of corpora is partitioned into documents. However, the problem with the direct use of documents is the big variance of document size. There are corpora that limit the size of their documents, e.g. documents in Brown corpus [10] contain 2000 words and then end on the sentence boundary even if it is inside the paragraph. On the other hand, corpora like Bank of English, based on the motto "More is better", throw out no text and therefore the size of documents can range from short newspaper notices to the whole books.

Taking into account the big variance and all the possible problems with topic shift within one document we decided to define the notion of context differently. We work with the context window $< -N, N >$, where N is the number of words on each side of the focus word. The context respects (not crosses) the document boundaries and ignores paragraph and sentence boundaries. The consequence of such definition is the symmetry of the relatedness measure.

2.3 Finding Collocations

Collocations of a given word are statements of the habitual or customary places of that word [11]. We have already mentioned the need of exclusion of collocations from our data to not contaminate clusters of semantically related words. We use the standard method of MI-score [12] to automatically identify the words that form a collocation. The only aspect of the process that is not routine is the extraction of three and more words collocations. It is implemented as a sequential process of piecewise formation of $n+1$-word collocation from possible n-word collocations. Considering the huge amount of data we are dealing with (100 million words corpora) it is obvious that the process of more-words collocations retrieving is time and resource demanding. (That is why we have used the capacity of a super-computer).

The side effect of collocation identification is the partial solution of the word sense ambiguity problem. As our method does not employ soft clustering (see below), the process is forced to decide to what cluster an ambiguous word will be adjoined. Applying collocation concept the word forming a collocation can belong to one cluster as a part of one particular collocation and to the other as a part of another collocation.

3 Arrangement of Experiments

We have been experimenting with two different corpora – large English corpus containing about 121 mil. of words and large Czech corpus containing about 120 mil. words (exact numbers can be found in Table 1). Data have been processd to exclude functional words using stop-list.

Table 1. Size of corpora used in experiments

# of	Czech	English
tokens	121,493,290	119,888,683
types	1,753,285	490,734
documents	329,977	4,124

The first step in the clustering process has been stemming or lemmatization (an assignment of the base form – lemma) of all word forms. The stemming algorithm for English can be implemented as a quite simple procedure which gives satisfactory results. On the other hand, lemmatization in the highly inflectional language like Czech needs carefully designed morphological analyzer. This effort is compensated by the reduction of items to be clustered (and therefore the time needed to process all data) and at the same time by the increase of occurrences of counted items and therefore by the increase of the statistical relevance of obtained data.

In order to eliminate singularities in statistics and to reduce the total number of the processed bigrams of words, we have restricted input data in several ways. The context of each word is taken as a window of 20 words on both sides. The minimal frequency of base forms has been set to 20 and the minimal frequency of bigrams to 5. Table 2 depicts exact values obtained from the Czech corpus.

Table 2. Statistics obtained from the Czech corpus

# of	
different lemmata	1,071,364
lemmata with frequency ≥ 5	218,956
lemmata with frequency ≥ 20	95,636
bigrams with frequency ≥ 5	25,009,524
lemmata in bigrams with frequency ≥ 5	72,311

The next task is to create lists of characteristic words for each context. The list of words sorted according to the decreasing MI score is prepared for each word. The MI score is used only to this ordering, in the following steps the particular values of the score are not taken into consideration. The size of such lists is limited to 500 words.

The calculation of distance between two words is motivated by the observation that semantically related words have similar characteristic lists. The difference of ranks for all the words from both lists is computed and 10 smallest differences are summed to form the distance. The graph in Figure 1 shows the relation between the number of bigrams and the computed distance.

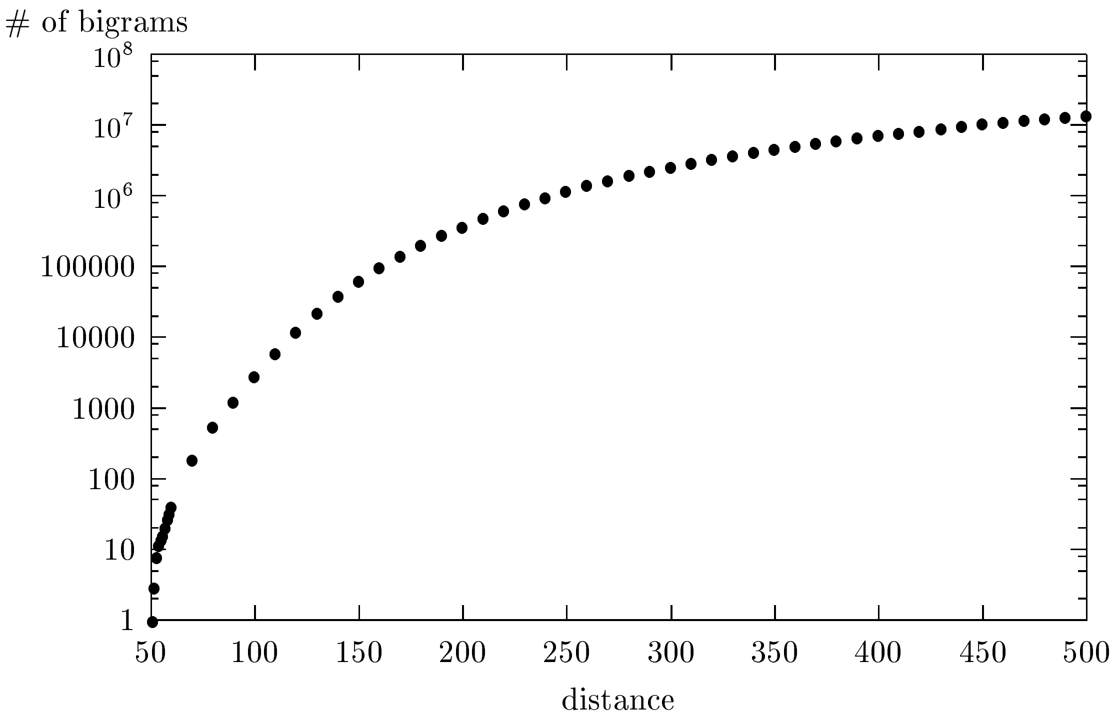

Fig. 1. Relation between the number of bigrams and their computed distance

4 Word Clustering

Data clustering, also known as unsupervised classification, is a generic label for a variety of procedures designed to find natural groupings (clusters) in multidimensional data, based on measured or perceived similarities among the patterns [13]. Cluster analysis is a very important and useful technique which forms an active research topic. Hundreds of clustering algorithms that have been proposed in the literature can be divided into two basic groups – partitional clustering and hierarchical clustering. Partitional algorithms attempt to obtain a partition which minimizes the within-cluster scatter or maximizes the between-cluster scatter. Hierarchical techniques organize the data in a nested sequence of groups that can be displayed in the form of a dendrogram (tree) [14].

Partitional clustering techniques are used more frequently than hierarchical techniques in pattern recognition. However, we argue that the number of clusters in the data, their shapes and sizes, depend highly on the particular application that should benefit from the clustered data. As our aim is to find clustering of the large vocabulary that could be used in many successive natural language tasks and for various application, the hierarchical techniques give more flexible outputs with universal (more general) usage. The weak point of this decision is the need of a heuristic to cut the dendrogram to form a partition required by a particular application.

The basic families of hierarchical clustering algorithms are single-link and complete-link algorithms. The former outputs a maximally connected subgraph, while the latter creates a maximally complete subgraph on the patterns. Complete-link clusters tend to be small and compact, on the other hand, single-link clusters easily chain together [14]. For our experiment, we have implemented a single-link clustering algorithm. The computational cost of this algorithm is acceptable even for the enormous number of words we are working with, contrary to the complete-link clustering algorithm that cannot directly benefit from the sorted list of distances and has to refresh the information about the distances each time a word is assigned to a cluster. The pseudo-code of the implemented algorithm can be seen in Figure 2.

As stated above, the hierarchical clustering algorithms output the dendrogram – a special type of tree depicting the iterative adjoining of words and clusters. A small subset of the dendrogram resulting from our experiments with corpus data can be found in Figure 3.

The final dendrogram has more than 40,000 nodes. As it is impossible to work with the whole tree in manual linguistic exploration of the results, we have implemented a simple procedure that, traversing the dendrogram, answers the question how a word is related to another one. Each particular line written by this procedure corresponds to the link of the dendrogram leading to a smallest partition of words covering both focus words. An example output of the procedure applied to words *exhalation* and *cleanup* is displayed in the following figure:

```
locateclust (id):
    Path ← ∅
    while clusters[id] not closed:
        Path ← Path ∪ {id}
        id ← clusters[id]
    foreach i ∈ Path:
        clusters[i] ← id
    return id

hierarchy():
    foreach < rank, id1, id2 >∈ sortbgr:
        c1 ← locateclust (id1)
        c2 ← locateclust (id2)
        if c1 ≠ c2:
            clusters[c2] ← c1
            hierarchy[c1] ← hierarchy[c1] ∪ {< c2, rank >}
            hierarchy[c2] ← hierarchy[c2] ∪ {< c1, 0 >}
    return hierarchy
```

Fig. 2. Pseudo code of implemented clustering algorithm

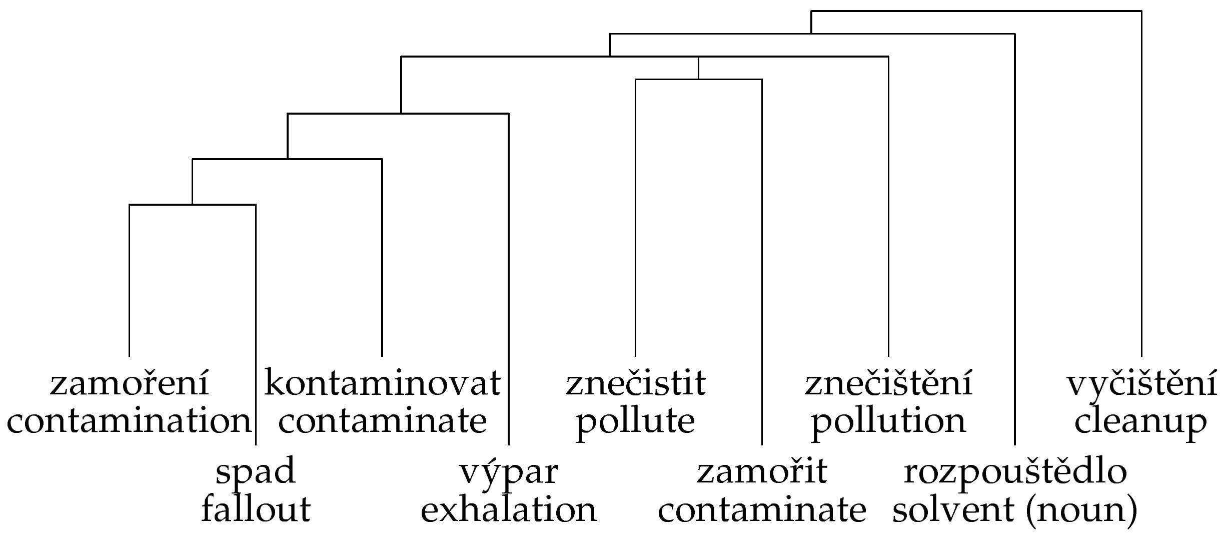

Fig. 3. An example of resulting dendrogram

výpar/exhalation
kontaminovat/contaminate (zamoření/contamination spad/fallout)
znečistit/pollute zamořit/contaminate
znečištění/pollution
rozpouštědlo/solvent(noun)
——
vyčištění/cleanup

Fig. 4. "Path" from word *výpar/exhalation* to *yčištění/cleanup*

5 Conclusions and Future Work

This paper presents a procedure of fully automatic finding of semantically related words. We have demonstrated that it is possible to work with large portions of text (100 million word corpora) and to find hierarchical partitioning of all reasonably frequent words. It is just this enormous size of the input corpus which is beside usually used methods that are applicable for toy-problems only. The amount of categorized words seems to be adequate for real applications, e.g. in the area of word sense disambiguation.

The automatic evaluation of the whole result set of 40,000 basic word forms is not possible today as there are no domain oriented dictionaries covering a significant portion of the Czech language. However, the comparison of the resulting clustering in three particular domains (weather, finance and cookery) is in good agreement with the human linguistic intuition.

The presence of polysemous and semantically ambiguous words poses the obstacle of any automatic word clustering. Our future effort will thus be focused on the correct treatment of these words. One of the possible solutions could be the incorporation of a mixture-decomposition clustering algorithm. This algorithm assumes that each classified pattern is drawn from one of the underlying populations (clusters) whose parameters are estimated from unlabelled data [15]. Mixture modeling allows soft membership that can be the answer to the semantic ambiguity problem.

Another direction for the future work will be oriented to objectivize the quality of clustering results. At present, the only way to asses the quality of the implemented procedure output is the manual checking of the results. We would like to employ the information from different sources like machine readable dictionaries, WordNet [16] and other semantic nets, and parallel corpora to purify the process of the evaluation from the subjectivity aspects.

References

1. Philip Stuart Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, 1993.

2. Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, 1994.
3. Steven Paul Finch. *Finding Structure in Language*. PhD thesis, University of Edinburgh, 1993.
4. C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge MA, 1999.
5. B. Boguraev and J. Pustejovsky, editors. *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge MA, 1995.
6. G. Grefenstette. *Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window-Based Approaches*, pages 205–216. MIT Press, Cambridge MA, 1996.
7. F. Smajda. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177, 1993.
8. M. P. Oakes. *Statistics for Corpus Linguistics*. Edinburgh University Press, 1997.
9. K. W. Church and W. A. Gale. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62, 1991.
10. W. N. Francis and H. Kučera. *Brown Corpus Manual*. Brown University, Providence, Rhode Island, revised and amplified edition, 1979.
11. F. R. Palmer. *Selected Papers of J. R. Firth 1952–1959*. London:Longman, 1968.
12. K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, March 1990.
13. A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall, 1988.
14. A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.
15. D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Willey and Sons, 1985.
16. G. A. Miller et al. Five papers on Wordnet. Technical report, 1993.

Publications in the FI MU Report Series are in general accessible via WWW and anonymous FTP:

`http://www.fi.muni.cz/informatics/reports/`
`ftp ftp.fi.muni.cz (cd pub/reports)`

Copies may be also obtained by contacting:

Faculty of Informatics
Masaryk University
Botanická 68a
602 00 Brno
Czech Republic