Query Expansion with an Automatically Generated Thesaurus *

José R. Pérez-Agüera and Lourdes Araujo
jose.aguera@fdi.ucm.es, lurdes@sip.ucm.es

Departamento de Sistemas Informáticos y Programación.
Universidad Complutense de Madrid. Spain.

Abstract. This paper describes a new method to automatically obtain a new thesaurus which exploits previously collected information. Our method relies on different resources, such as a text collection, a set of source thesauri and other linguistic resources. We have applied different techniques in the different phases of the process. By applying indexing techniques, the text collection provides the set of *initial terms* of interest for the new thesaurus. Then, these terms are searched in the source thesauri, providing the initial structure of the new thesaurus. Finally, the new thesaurus is enriched by searching for new relationships among its terms. These relationships are first detected using similarity measures and then are characterized with a type (equivalence, hierarchy or associativity) by using different linguistic resources. We have based the system evaluation on the results obtained with and without the thesaurus in an information retrieval task proposed by the Cross-Language Evaluation Forum (CLEF). The results of these experiments have revealed a clear improvement of the performance.

1 Introduction

A thesaurus is a structured list of terms, usually related to a particular domain of knowledge. Thesauri are used to standardize terminology and to provide alternative and preferred terms for any application. They are specially useful in indexing and retrieving information processes, by providing the different forms which a concept can adopt.

The three basic relationships between thesaurus terms are equivalence, hierarchy and associativity. Terms related by the equivalence relationship have an equivalent meaning, in different senses (they are synonyms, one is the translation to the other, its archaic form, etc). In a set of equivalent terms, one of them, distinguished as the *preferred* one, is the one used in the hierarchies and for indexing. Preferred terms are arranged into hierarchies with different numbers of levels. These levels go from the broadest type of term to the narrowest and most specific one. Finally, there can be associative relationships between terms which are not connected by a hierarchy, for example because they are narrower terms of different broad terms, but they still present some kind of relationship.

* Supported by project TIC2003-09481-C04

In spite of the great interest thesaurus have reached nowadays for web applications, most of them are manually generated, what is very expensive and limits its availability to some particular topics. Furthermore, a thesaurus usually requires to be periodically updated to include new terminology, in particular in modern terms, such as those related to computer science. These reasons make the automatic generation of thesauri an interesting area of research which is attracting a lot of interest. Research on automatic thesaurus generation for information retrieval began with Sparck Jones's works on automatic term classification [5], G. Salton's work on automatic thesaurus construction and query expansion [9], and Van Rijsbergen's work on term co-occurrence [10]. In the nineties Qui and Frei [7] [6] [8] worked on a term-vs-term similarity matrix based on how the terms of the collection are indexed. Recently, Zazo, Berrocal, Figuerola and Rodríguez [1] have developed a work using similarity thesauri for Spanish documents. Jing and Croft [4] have carried out an approach to automatically construct collection-dependent association thesauri using large full-text documents collections. Those approaches obtain promising results when applied to improve information retrieval processes.

This paper proposes to apply a combination of techniques to automatically obtain a new thesaurus for a particular knowledge domain. The method relies on different resources from which we extract selected information, thus taking advantage of the information previously gathered and processed, what improve both the accuracy and the efficiency. The domain of knowledge to which the new thesaurus is devoted, is characterized by a set of terms extracted from a document collection about the intended topic. This is done by applying indexing techniques. Then we use the information previously collected in other thesauri about these terms to construct the initial structure of the new one. Finally, the new thesaurus is enriched by searching for new relationships among its terms. These relationships are first detected using co-occurrence measures and then its type (equivalence, hierarchy or associativity) is characterized by using different linguistic resources, such as a dictionary and a POS tagger. Apart from the specific methods that we have applied, our proposal differs from previous works in that the relationship type (equivalence, hierarchy or associativity) of the generated thesaurus is identified. The system has been evaluated by comparing the results obtained in an information retrieval task, for which the expected results are perfectly defined, when a set of query terms are directly consulted, and when they are previously expanded with the generated thesaurus.

The rest of the paper proceeds as follows: section 2 describes the general scheme of the system, presenting their different phases and tools; sections 3 describes the enrichment of the relationships between the thesaurus terms and the identification of its type; section 4 presents and discusses the experimental results, and section 5 summarizes the main conclusions of this work.

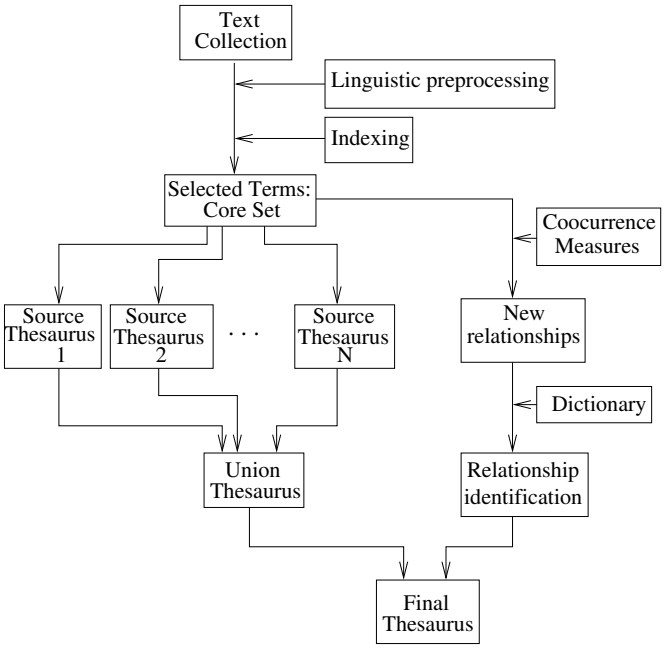

Fig. 1. Process to generate a new thesaurus.

2 System Overview

We combine different techniques to obtain a new thesaurus for a particular domain of knowledge. Figure 1 shows a scheme of the process. First of all, we perform a selection of terms, called the *core set*, from a text collection concerning the intended thesaurus domain. In this phase we apply linguistic pre-processing which consists of a POS tagging which allows selecting only those words of noun category, stemming, and elimination of stopwords. We apply TF-IDF to the candidate words in order to obtain the initial list of thesaurus terms.

The next step of the process is the generation of the *union thesaurus* from a set of source thesauri. The source thesauri that we have used are the following ones:

- EUROVOC, which contains concepts on the activity of the European Union.
- SPINES, a controlled and structured vocabulary for information processing in the field of science and technology for development.
- ISOC, thesaurus aimed at the treatment of information on economy.

Terms which appear in both, the core set and any source thesauri, are the term list of the *union* thesaurus. Furthermore, the relationships among the terms included in the new thesaurus are provided by the source thesauri. Figure 2 shows an example of generation of the *union* thesaurus. When the term *terremoto*

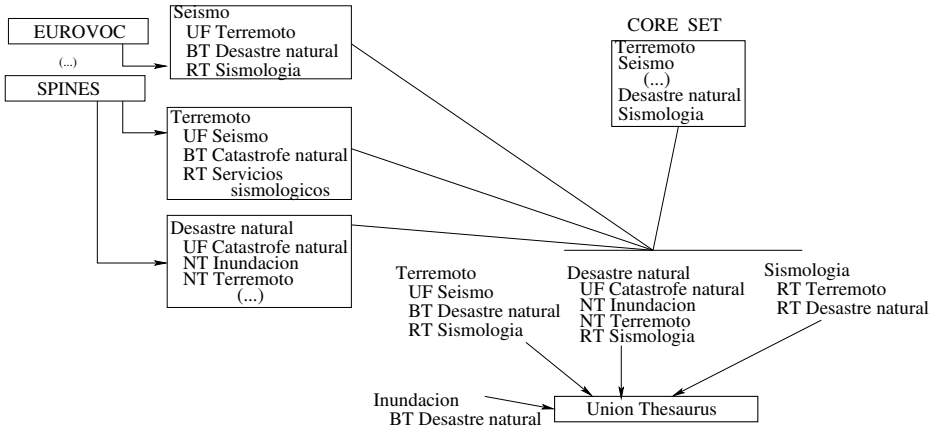

Fig. 2. Example of generation of the *union* thesaurus. UF stands for *used for*, NT for *narrower term*, BT for *broader term* and RT for *related term*.

(earthquake), which belongs to the core set, is searched in the source thesauri two entries are found, one in SPINES and the other one in EUROVOC. In EUROVOC *terremoto* (earthquake) belongs to an entry whose preferred term is *seísmo* (seism) and which also contains *desastre natural* (natural disaster) (BT), and *sismología* (seismology) (RT). In SPINES *terremoto* is the preferred term of an entry which also contains the synonym *seísmo*, the broader term *catástrofe natural* (natural catastrophe). In SPINES, *terremoto* also appears in other entry whose preferred term is *desastre natural*, and which also contains the synonym *catástrofe natural*, and the narrower terms *inundación* (flood) and *terremoto*. Accordingly, the *union* thesaurus presents entries whose preferred terms are *terremoto*, *desastre natural*, *sismología* and *inundación*, the terms of the core set.

The *union* thesaurus just described is now extended by detecting new semantic relationships among the terms of the set compose of the core set plus the terms taken from the source thesauri.

3 Enriching the Hierarchies

If a couple of terms to be related, appears in some of the source thesauri, this indicates the kind of its relationships. If they do not appear in the source thesauri, its possible relationship has to be investigated. The first step is to detect any kind of relationship, and then, in a second step the type of the detected relationship is identify.

For the extraction of semantic relations between terms we have chosen the statistical method of the Vector Space Model [3]. To apply the vectorial model we defined a vector of features for each term from the documents in which it

appears. The values of this vector are estimated by counting the co-occurrences of the terms in the documents. After testing different classic measures, such as Dice, Jaccard and Cosine, we have chosen Cosine, which provides the best results for our work.

Once we have determined the pairs of terms for which the semantic similarity is significant enough (the similarity is above a threshold value of 0.3), we have to determine the type of the relationship between these pairs of terms: equivalence, hierarchy or associativity. We assume that the degree of semantic similarity between a term and a preferred one depends on the type of the relationship between them: equivalent terms have the highest values, followed by terms which belong to the same hierarchy, and related terms have the lowest values. Accordingly, we concentrate in detecting the hierarchical relationships, and the higher and lower values of semantic similarity of these pairs are considered respectively as the top and the bottom threshold values of this type of relationship. Then term pairs with sematic similarity over the top threshold are assigned the equivalence relationship, while terms with a value of semantic similarity below the bottom threshold are considered related terms.

Let us now to consider the technique used to detect hierarchical relationships. It relies on the assumption that in a dictionary the entries for a term which is an instance of a more general concept contain a reference to the term for this general concept. Furthermore, we assume that the references to more general terms usually adopt some predefined structures. We have considered the following set of structures for the detection of hierarchical relationships:

noun
noun adjective
noun noun
noun preposition noun
noun preposition article noun

We have developed our experiments in Spanish, using the RAE (Real Academia Española) dictionary, applying a part-of-speech (POS) tagging of the dictionary entries in order to detect the selected structures. For query expansion we use the method proposed by Qiu y Frei [7], which selects expansion terms according to their similarity with all query terms.

4 Experiments and Results

The prototype developed for our experiments has been implemented using the programming language Java. This prototype has been run on a computer Intel Pentium IV Hyper-Threading 3.40 GHz, with 2GB of RAM memory.

In order to provide a quantitative measure for the quality of the generated thesaurus, we have decided to evaluate its usefulness when it is applied to an information retrieval task. Specifically, we used the thesaurus to perform a term-to-term query expansion, i.e. for identifying terms related with the query terms in order to improve the retrieval capability.

With the aim at being as fair as possible, in the selection of tests we have taken a set of tests used in the CLEF (Cross-Language Evaluation Forum) for the Spanish language. The collection and tests used come from EFE94.

For the evaluation of the system we have used *trec_eval* package, with the measures of precision and recall [2]. Recall is the fraction of the relevant documents which have been retrieved and precision is the fraction of the retrieved documents which are relevant. Besides, we use R-precision, which is the precision after retrieving R documents, where R is the total number of relevant documents for the query. As test set we have used a total of 50 queries extracted from the batteries provided by CLEF in 2001.

Table 1 shows the results obtained performing query expansion with different thesauri. We can observe that the results for the automatically generated thesaurus (last row) are significantly better than those obtained with the source and *union* thesauri, since we obtain a general improvement of 9,47% in the precision and of 9.99% in the recall, while the source and *union* thesauri obtain negative results. On the one hand, the query expansion achieved by using the thesaurus enlarges the set of search terms and thus recall improves. On the other hand, precision also improves because the percentage of relevant documents retrieved with the query expansion is larger than the percentage for the original query. Because the Qui & Frei expansion method that we apply for the query expansion requires similarity measures to work appropriately, the source and *union* thesauri, which do not provide such measures, do not achieve any improvement in the retrieval. We have also tested a direct expansion method, which does not use similarity measures, and it also provides much worse results for the source and *union* thesauri.

We have performed a number of experiments in order to determine the influence of the different steps of the linguistic preprocessing on the results. Table 2 shows the results with and without POS tagging. The first row shows the results without query expansion. The other two rows present the results expanding with the thesaurus generated without (second row) and with POS tagging (third row). We can observe that the POS tagging not only improve the different measures (precision, R-precision and recall) but also reduces the index size, what leads to a significative decrease of the execution time.

Query	Precision	Recall	Improve. Precision	Improve. Recall
Baseline	0.4460	0.7584	–	–
Spines	0.3624	0.6416	- 18.74%	- 15.40%
Eurovoc	0.3730	0.6550	- 16.37%	- 13.63%
ISOC-Economy	0.3728	0.6415	- 16.41%	- 15.41%
Union Thesaurus	0.3727	0.6556	-16.43%	- 13.55%
Final Thesaurus	0.4927	0.8426	+ 9,47%	+ 9.99%

Table 1. Comparison of results obtained applying query expansion with different thesauri.

Query	Precision	R-Precision	Recall	Index size	Time
Baseline	0.4460	0.4482	0.7584	352.534	-
Thesaurus	0.4789 (+ 6.86%)	0.4745 (+ 5.54%)	0.8460 (+ 10.35%)	352.534	242
Th. w. POS	0.4906 (+ 9.09%)	0.4886 (+ 8.71%)	0.8454 (+ 10.29%)	321.612	220

Table 2. Comparison of results obtained applying POS tagging in the thesaurus generation. Th. w. POS stands for Thesaurus with POS tagging. Time is given in minutes.

Table 3 shows the results with and without stemming. As in the case of POS tagging, stemming not only improves the different measures but also reduces the index size, and thus the execution time.

Table 4 compares the results when *specific stopwords* are eliminated of the core set. Specific stopwords are not typical stopwords, but they are words too frequent in the collection to be good discriminators for thesaurus construction. Examples of specific stopwords are months, name of the days, etc. The default thesaurus has been generated with POS tagging and stemming. In this case the improvement of the measures is smaller than in the other cases. We think that it is because the frequency must not be the only factor to take into account to determine the specific stopwords, but the degree of relationship with other words of the intended domain must also be considered.

5 Conclusions and Future Works

This paper shows how to use handmade thesauri for the automatic generation of new thesauri. There exists a large amount of handmade thesauri, which are very useful as knowledge bases for the automatic generation of thesauri[1]. Furthermore, we have defined a methodology to combine linguistic methods and statistical methods for the automatic generation of thesauri. This is one of the ways in which natural language processing can improve the performance of information retrieval processes. Results have shown the usefulness of the generated thesaurus, improving both, recall and precision measures in an information retrieval task. Recall improves because the list of search terms is enlarged with the query expansion. And precision also improves since most of the new documents

[1] Web Thesaurus Compendium: http://www.ipsi.fraunhofer.de/ lutes/thesoecd.html

Query	Precision	R-Precision	Recall	Index size	Time
Baseline	0.4460	0.4482	0.7584	352.534	-
Thesaurus	0.4789 (+ 6.86%)	0.4745 (+ 5.54%)	0.8460 (+ 10.35%)	352.534	242
Th. w. Stem.	0.4832 (+ 7.69%)	0.4815 (+ 6.91%)	0.8428 (+ 10.01%)	315.457	216

Table 3. Comparison of results obtained applying stemming in the thesaurus generation. Th. w. Stem. stands for Thesaurus with stemming. Time is given in minutes.

Query	Precision	R-Precision	Recall	Index size	Time
Baseline	0.4460	0.4482	0.7584	352.534	-
Thesaurus	0.4906 (+ 9.09%)	0.4906 (+ 9.09%)	0.4906 (+ 9.09%)	321.612	220
Th. wo. SSW	0.4927 (+ 9.47%)	0.4916 (+ 9.27%)	0.8426 (+ 9.99%)	321.503	220

Table 4. Comparison of results obtained by eliminating specific stopwords in the thesaurus generation. Th. wo. SSW stands for Thesaurus without specific stopwords. Time is given in minutes.

added to the retrieved list are relevant, and thus the rate of relevance improves. We have also shown the advantages of some preprocessing steps, such as POS-tagging and stemming, used in the process of selection of the term list which characterizes the intended domain.

Given the promising results obtained, we plan to investigate how to improve the different phases of this process. In particular, we plan to apply a more exhaustive linguistic analysis for the identification of semantic relationships, as well as using Wordnet as another source of information for the thesaurus generation.

References

1. Angel F. Zazo and Carlos G. Figuerola and Jose L. Alonso Berrocal and Emilio Rodríguez. Reformulation of queries using similarity thesauri. *Information Processing and Management*, 41(5):1163–1173, 2005.
2. Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
3. G. Salton. *Automatic Information Organization and Retrieval*. McGraw Hill Book Co, 1968.
4. Y. Jing and W. Bruce Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 146–160, New York, US, 1994.
5. K. Sparck Jones and R.M. Needham. Automatic Term Classification and Retrieval. *Information Processing and Management*, 4(1):91–100, 1968.
6. Yonggang Qiu and Hans-Peter Frei. Applying a similarity thesaurus to a large collection for information retrieval, 1993.
7. Yonggang Qiu and Hans-Peter Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, US, 1993.
8. Yonggang Qiu and Hans-Peter Frei. Improving the retrieval effectiveness by a similarity thesaurus. Technical Report 225, Dept of Computer Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, 1995.
9. G. Salton, C. Buckley, and C. T. Yu. An evaluation of term dependence models in information retrieval. In *SIGIR '82: Proceedings of the 5th annual ACM conference on Research and development in information retrieval*, pages 151–173, New York, NY, USA, 1982. Springer-Verlag New York, Inc.
10. C.J van. Rijsbergen, D.J. Harper, and M.F. Porter. The selection of good search terms. *Information Processing and Management*, 17(2):77–91, 1981.