Term Expansion Using Stemming and Thesauri in Spanish

Ángel F. Zazo, Carlos G. Figuerola, José Luis A. Berrocal and Emilio Rodríguez
Grupo de Recuperación de Información
Departamento de Informática y Automática - Universidad de Salamanca
37008 Salamanca - SPAIN
{*afzazo, figue, berrocal, aldana*}*@usal.es*

Abstract

The objective of our participation this year in the Spanish monolingual task at CLEF2002 is to continue the study in term expansion. Last year we showed results in stemming. Now, our effort is centered in term expansion using thesauri. Many words that derive from the same stem have a close semantic content. However other words with very different stem origin have near semantic sense. In this case, the analysis of the word relationships in document collection can be used to construct a thesaurus of related terms. After, the thesaurus is used to expand a term with the best related terms.

1 Introduction

A major problem in word based information retrieval (IR) is *word-mismatch or vocabulary problem* [5]. Lexical figures as the synonymy and polysemy cause that the same concept can be expressed with different words and the same word can appear in documents that treat about different topics. The performance of these systems depends on the number of query terms. The problem is less severe for long queries because more index terms are included and there is more chance to appear in relevant document. Short queries are poor for recall and precision. They do not take into account for the variety of words used to describe a topic, and are too broad to retrieve relevant documents on specific topics. Our interest is centered in queries with very few terms. They have special importance in Web search engines, typically with one to three terms by query [24].

Many techniques have been used to try reduce this problem, one of them is query expansion. Query expansion methods have been investigated for almost as long as the study of information retrieval. This technique involves two basic steps: expanding the original query with new terms, and reweighting the terms in the expanded query. With query expansion the retrieval performance can improve, although, in contrast the computational cost or the response time can increase. The techniques developed can be classified as user-assisted or automatic. Automatic query expansion techniques require no effort on the part of the user and have a significant advantage over manual techniques such as relevance feedback [18] and manual thesauri.

In order to expand the query, words or phrases with similar meaning to those of the initial query must be used. Several approaches exist to carry out this task, the more important is the use of a thesaurus. A thesaurus is a classification system composed of word or phrases and for each, a set of related words. In information retrieval, thesauri are used for helping with the query formation process. Likewise, stemming can be thought of as a mechanism for query expansion, and it can be seen similar to using a thesaurus. Some stemmers can be created or modified using same techniques as for thesaurus construction [25].

This paper explores stemming and thesaurus approach (association and similarity thesauri) to term expansion. We assume the well-known vector space model, but queries are first expanded to help improve the retrieval performance.

2 Stemming

Query expansion through relational morphology behaves as a stemming procedure, and tries to generate all related word forms from the query words. The concept of term (really *index term*) is not exactly

the same as word. After removing stops words, which cannot be considered terms as such, we have the case of words derived from the same stem, which can be attributed a very close semantic content [9]. Derivatives, inflections, alteration in gender and number, etc., make it advisable to group these variants under one term. For instance, the Spanish term *carne* (meat) is expanded to: *carne, carnes*; *carnicero, carniceros, carnicera, carniceras* (butcher); *carnicería, carnicerías* (butcher's shop); *carnal, carnales* (carnal); *cárnica, cárnico, cárnicas, cárnicos* (related to meat); *carnaza, carnazas, carnada, carnadas* (feed, bait); *carnoso, carnosas, carnoso, carnosos* (beefy), etc.

Stemming is usually viewed as a recall-enhancing method, but it can sometimes improve precision at low recall levels. However, the effectiveness of stemming has been the object of certain discussion. Harman [8], after trying several algorithms (for English) concluded that none of them increase effectiveness in retrieval. Popovic and Willett [14] argue that the effectiveness of a stemmer is a function of the morphological complexity of the language in the document set. Krovetz [11] and Hull [9] found that stemming slightly improves recall and even precision in some collections (e.g., when documents and queries are very short).

As regards Spanish, diverse stemming algorithms were applied in some TREC and CLEF conferences. Generally, they use the same base algorithms as for English, but with suffixes list and rules for Spanish. This adaptation is quite poor [2]. On the other hand, the use of n-grams has been proposed to obviate the problem [16]. In prior works, however, we were able to verify the scarce effectiveness for information retrieval, as well as the inadequacy of the well-known Porter algorithm for languages such as Spanish [3].

Last year at CLEF2001, we use a finite states machine stemmer, which represents the modifications undergone by a stem when certain suffix is attached to it [4]. For each suffix a set of rules exists to drive how that suffix is attached into the stem. For one suffix there may be a large number of variants and exceptions, therefore the resulting automaton can be quite complex. In order to stem a word, the longest suffix agreeing with the end of this word is searched, and the respective automaton is formed with the rules for that suffix. The network of this automaton is searched with the word to be stemmed and the chains obtained in the terminal node are contrasted with a dictionary of stems. If the chain obtained is found in the dictionary, the stem is considered to be correct.

We can distinguish between two classes of stemming: inflectional and derivational. Whereas the former describes predictable changes a word undergoes as a result of syntax, and they have no effect on a word's part-of-speech, in contrast, the latter may or may not affect a word's part-of-speech, and may or may not affect its meaning [11]. In general, there is a few semantic distance between two inflection of the same stem (*libro* and *libros*). On the contrary, the meaning of derivatives can be very different, e.g., *sombra* (shadow), *sombrilla* (parasol, sunshade), *sombrero* (hat).

The impact of our stemmers for the Spanish monolingual track at CLEF2001 was presented in [4]. For all query fields (title + narrative + descriptive), the improvement is only 3% for inflectional stemming over unstemming. Derivational stemming is even a little bit worse than no stemming.

This year we have corrected small bugs on stemmers, and only inflectional stemming was applied. The objective is to see the improvement, taking into account only the `ES-title` field of the queries. After, we compare the results with those of applying thesaurus.

3 Thesaurus

One of most important methods for query expansion is the application of a thesaurus. A general thesaurus could be used, nevertheless, this usually does not give good results (e.g. [23]). The relations in a general thesaurus are usually not valid in the scope of the document collection. Better results are obtained if thesauri, or other expansion techniques, are constructed from the document collection. When the thesaurus is constructed automatically, without additional user relevance information, several approaches are distinguished [7]:

- Automatic term classification (term co-occurrences statistics) [12]. The similarity between terms is made on the basis of the Association Hypothesis: *if a term is good at discriminating relevant from non-relevant documents, any closely associated term is also likely to be good at this* [22, p. 104].
- Use of document classification [1]. Documents are first classified and infrequent terms found in a document class are used to construct the thesaurus of related terms.
- Concept based query expansion [19, 15]. A similarity thesaurus is constructed making the transposition of the document-terms matrix: the documents constitute the indexing features of the terms.

- Phase-finder expansion [10]. A phrase is a sequence of terms whose part-of-speech satisfy one specific rule of the phrases detector (NNN, NN, AN, etc.). Phrases in the text are identified and they are associated to terms.
- Based on syntactic information [6]. The terms relations are generated on the basis of linguistic knowledge and co-occurrences statistic. Grammars and dictionaries are used to obtain the list of related terms for each term.

We have experienced approaches 1 and 3, because they are relatively simple and effective.

A thesaurus is a matrix that measures term relations [20]. This matrix is used to expand the query terms with related terms. The matrix can be seen as a semantic description of terms, which reflects the influences of terms in the conceptual descriptions of other terms.

We note two fundamental aspects to apply the matrix in our experiments. First, the expansion is made with *better* related terms. No threshold values are taking into consideration: the terms with high related values are selected for each original query term. Second, the entire query, i.e., the query concept [15], is taken into account. The top ranked terms for the entire query are considered.

Before to go on, we must say that results on term expansion using thesauri differ. Several works cited previously show not bad results. In the other hand, for example, [13] offers perhaps the most critical study of term co-occurrence based models. Earlier studies [21] even showed better results with randomly selected terms than those from term co-occurrence statistic. Despite, we have obtained satisfactory results, perhaps due two notes previously mentioned.

3.1 Association matrix

Term co-occurrence has been used frequently in IR to identify some of the semantic relationships that exist between terms. In fact, this idea is based on the Association Hypothesis [22, p. 104]. Query terms are good to distinguish relevant from non relevant documents, then their associated terms will be good as well, and can be added to the original query.

Several coefficients has been used to calculate the degree of relationship between two terms. All of them measure the number of documents in which they separately occur, in comparison with the number of documents in which they co-occur. In our experiments three well-known coefficients have been calculated: Tanimoto, Cosine and Dice [19], but, due to limits in the number of runs to submit, only results of the first was sending to evaluation.

$$\text{Tanimoto}(t_i, t_j) = \frac{c_{ij}}{c_i + c_j - c_{ij}} \qquad \text{Cosine}(t_i, t_j) = \frac{c_{ij}}{\sqrt{c_i \cdot c_j}} \qquad \text{Dice}(t_i, t_j) = \frac{2 \cdot c_{ij}}{c_i + c_j}$$

where c_i and c_j are the number of documents which terms t_i and t_j occur, respectively, and c_{ij} is the number of documents which t_i and t_j co-occur. The coefficients have values between 0 and 1: if two terms occur exclusively in the same documents, the associated coefficient is 1, if there isn't a document which they co-occur the value is 0. The association matrix is symmetric and the diagonal elements are equal to 1.

3.2 Similarity matrix

The similarity matrix measures term-term similarities, instead of term-term co-occurrences. To compute the element values each term is indexing by the documents in which it occur, i.e., the roles of terms and documents are interchanged. A complete paper explains this theory is [15]. The broad outlines are: first, each term in the document vector space is represented by a vector, which elements are computed adapting the normalized *tf·idf* weighting scheme to this new situation. We use the same calculus as in [15]. Second, to compute the similarity between two terms the simple scalar vector product is used. We also use it.

Computation for every terms produces the similarity matrix. It is a symmetric matrix, with diagonal elements equal to 1. All values are between 0 and 1.

3.3 Expansion of the query

The objective to use the association or similarity matrix is expand the entire query, not only separate terms. One term can be included in the list of expanded terms if it has a high relational value with all

Collection	EFE
Documents	215.738 (513 MB raw)
Queries	50
Total index terms	352.777
Averaged doc length (words)	333,68 (max. 2210, min. 9)
Averaged doc length (unique index terms)	120,48
Averaged query length (unique index terms)	2.62 (`ES-title`) 20,48 (all)

Table 1: Collection.

query terms. To obtain the better terms to expand, each term of the query is expanded with all related terms. A new value is computed making the product of the weight of the query term by the corresponding association/similarity value of related terms. For all query terms, all values are added for each candidate term to expand. This values is the relational value with entire query. Only top ranked terms are used to expand the original query.

Finally, it necessary to calculate the weight of the term added to the query. Is clear it depends on the relational value with entire query above mentioned. In [15] it is used the sum of the weight of original query to reduce this value. We use it as well. Nevertheless, in others experiments we use the number of original query terms. In this way, the results are not conditioned with vector normalization of the query.

4 Experiments

The 50 queries of the CLEF2002 Spanish monolingual collection were executed in four modalities:

- without stemming for all query fields, `ES-title`, `ES-desc` and `ES-narr` (usalNNTDN)
- applying inflectional stemming for all query fields (usalFNTDN)
- without stemming using the `ES-title` query field and applying query expansion based on association thesaurus with 300 terms added (usalNAT)
- without stemming using the `ES-title` query field and applying query expansion based on similarity thesaurus with 300 terms added (usalNST)

The first two runs was submitted to compare the results with 2001 campaign. The second two was submitted to check the thesaurus expansion experiments. Table 1 shows the collection used for experiments. Only `TITLE` and `TEXT` fields of documents are used. For queries, the table indicates the averaged number of unique index terms for `ES-title` field and for all fields.

For our experiments, we converted all words to lowercase and accentuated vocals to unaccentuated ones, and included number as index terms. The number of stop words was 573. We used the well-known *tf·idf* scheme and recommendations in [17], and the simple scalar vector product to calculated the similarity between query and documents. Only the document vectors were normalized.

4.1 Results in stemming

The result of the first two official runs is showed in Fig. 1. The improvement respect last year is similar. Table 2 indicates the improvement, averaged over all queries, in average precision non-interpolated, averaged R-Precision and averaged precision at 10 documents seen. We include the last measurement because ten documents at once are habitual to show in the user interface.

Measurement	No stemming (usalNNTDN)	Inflectional stemming (usalFNTDN)	Improvement
avg. precision	0.3908	0.4051	3.66%
avg. R-Precision	0.3844	0.4076	6.04%
avg. prec. at 10 docs	0.4840	0.4900	1.24%

Table 2: Improvement using inflectional stemming (all query fields).

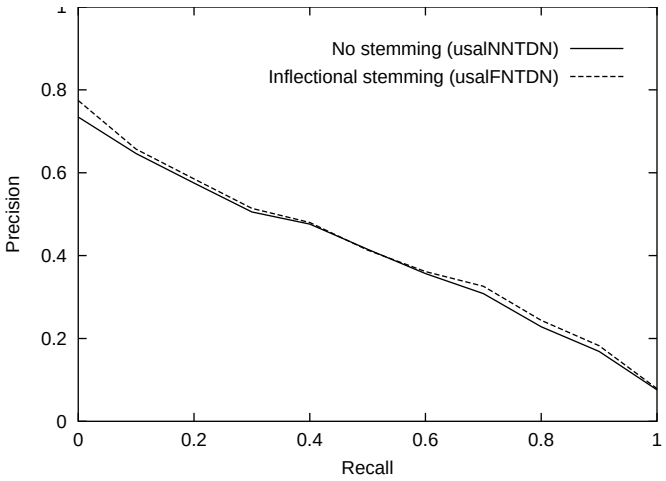

Figure 1: Results of official runs.

4.2 Results in query expansion

We compute the association and similarity thesauri from the document collection. The objective is to obtain the improvement taking into account only the `ES-title` field of queries. For efficiency sake, only the terms in original queries (really, terms in `ES-title` field) were selected as entries in thesauri. We don't use word or text windows, as habitual in other studies: whole document (`TEXT` and `TITLE` fields) is treated as an unique word window. We also do not apply stemming. In our experiments in query expansion using association thesaurus, only results with Tanimoto coefficient has been sending to evaluation. The number of terms added to expand the original query was 300 in both association and similarity. This is a reasonable number [15]. For comparison purposes, we have calculate recall-precision values for only `ES-title` query field without applying expansion or stemming. After, we compare the results with those of applying only inflectional stemming.

Figure 2 shows precision-recall curves. Table 3 shows the improvement with respect to no stemming.

5 Conclusions and future work

In this work we explore stemming and thesaurus approach (association and similarity thesauri) to query or term expansion: queries are first expanded to help improve the retrieval performance. The results show that these techniques are valid. The first experiment was carry out over all query fields applying inflectional stemming. The improvement is 3.66 % in averaged precision over unstemming. Taking only the `ES-Title` field the improvement is similar, 4.39%. Our interest is centered in queries with very few terms. They have special importance in Web search engines, typically with one to three terms by query. In our query collection the averaged query length is 2.62 for `ES-title`.

Betters results are reached with automatic thesauri expansion. Thesauri are constructed from the document collection. For query expansion using association thesaurus (Tanimoto coefficient) the improvement is 9.97%, and 20.05% for similarity thesaurus. Is important to show the improvement in averaged precision at 10 documents seen, given that usually it is the number of documents showed at once in the user interface. The improvement is 6.63% and 18.67%, respectively. This values indicate that query

	No expansion	Thesauri expansion		Inflectional
Measurement	no stemming	Tanimoto (usalNAT)	Similarity (usalNST)	Stemming
avg. precision	0.2618	0.2879 (9.97%)	0.3143 (20.05%)	0.2733 (4.39%)
avg. R-Precision	0.2752	0.2952 (7.27%)	0.3185 (15.73%)	0.2866 (4.14%)
avg. prec. at 10 docs	0.3320	0.3540 (6.63%)	0.3940 (18.67%)	0.3460 (4.21%)

Table 3: Improvement (`ES-Title` field).

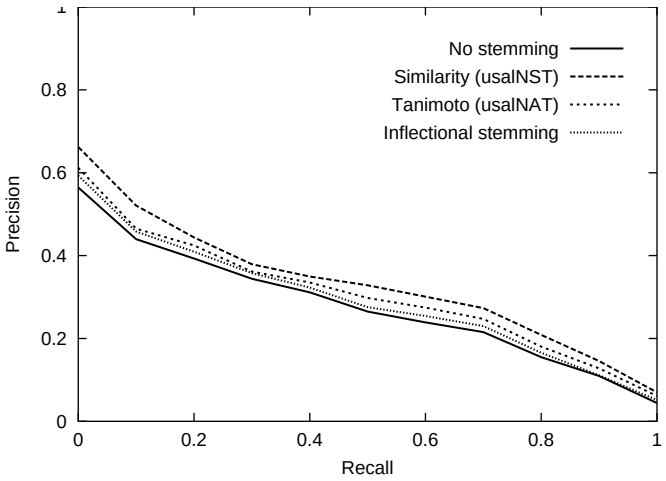

Figure 2: Precision-recall curves for `ES-title` field.

expansion using thesauri is a valid mechanism to improve the retrieval performance. Note, however, that experiments was carry out with 300 terms added to original query. This is the disadvantage: the computational cost increase with the number of query term.

For the future, besides to correct some bugs in the stemmer, we need to extend the experiments in query expansion: test different co-occurrence coefficients, test different weighting schemes for similarity thesaurus construction, test different weighting schemes for expanded terms added to query, etc. Other experiment can be carry out to increase the accuracy of stemmers using thesauri to determine semantic relations between terms.

References

[1] C. J. Crouch and B. Yang. Experiments in automatic statistical thesaurus construction. In N. J. Belkin, P. Ingwersen, and A. M. Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*, pages 77–88. ACM, 1992.

[2] C. G. Figuerola. La investigación sobre recuperación de información en español. In C. Gonzalo García and V. García Yedra, editors, *Documentación, Terminología y Traducción*, pages 73–82. Síntesis, Madrid, 2000.

[3] C. G. Figuerola, R. Gómez Díaz, and E. López de San Román. Stemming and n-grams in Spanish: an evaluation of their impact on information retrieval. *Journal of Information Science*, 26(6):461–467, 2000.

[4] C. G. Figuerola, R. Gómez Díaz, A. F. Zazo Rodríguez, and J. L. Alonso Berrocal. Spanish monolingual track: the impact of stemming on retrieval. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Darmstadt, Germany, September 2001, Revised papers*, pages 253–261. Springer-Verlag, Berlin, etc., 2001. (Lecture Notes in Computer Science; Vol. LNCS 2406).

[5] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Comunications of the ACM*, 30(11):964–971, November 1987.

[6] G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In N. J. Belkin, P. Ingwersen, and A. M. Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*, pages 89–97. ACM, 1992.

[7] C. Han, H. Fujii, and W. Croft. Automatic query expansion for japanese text retrieval. Technical Report UM-CS-1995-011, Department of Computer Science, Lederle Graduate Research Center, Univer-

sity of Massachusetts, 1995. En línea: ftp://ftp.cs.umass.edu/pub/techrept/techreport/1995/UM-CS-1995-011.ps.

[8] D. K. Harman. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1):7–15, 1991.

[9] D. A. Hull. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84, 1996.

[10] Y. Jing and W. B. 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.

[11] R. Krovetz. Viewing Morphology as an Inference Process,. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–203, Pittsburgh (US), 1993.

[12] J. Minker, G. G.A. Wilson, and B. Zimmerman. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8(6):329–348, 1972.

[13] H. J. Peat and P. Willet. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

[14] M. Popovic and P. Willet. The effectiveness of stemming for natural language access to slovene textual data. *Journal of the American Society for Information Science*, 43(5):384–390, 1992.

[15] Y. Qiu and H.-P. Frei. Concept-based query expansion. In *Proceedings of the Sixteenth International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh (US), 1993.

[16] A. M. Robertson and P. Willett. Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1):48–67, 1998.

[17] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[18] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[19] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New-York, 1983.

[20] H. Schutze. Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787–796, Minneapolis, 1992.

[21] A. Smeaton and C. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[22] C. van Rijsbergen. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979. También en línea: `http://www.dcs.gla.ac.uk/Keith/`.

[23] E. Voorhees. Query expansion ussing lexical-semantic relations. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, 1994.

[24] D. Wolfram, A. Spink, B. J. Janses, and T. Saracevic. Vox populi: The public searching of the web. *Journal of the American Society for Information Science and Technology*, 52(12):1073–1074, 2001.

[25] J. Xu and W. B. Croft. Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems*, 16(1):61–81, 1998.