Towards the development of heuristics for automatic query expansion

Jesús Vilares, Manuel Vilares, and Miguel A. Alonso

Departamento de Computación
Facultad de Informática, Universidad de La Coruña
Campus de Elviña s/n, 15071 La Coruña, Spain
jvilares@mail2.udc.es, vilares@dc.fi.udc.es, alonso@dc.fi.udc.es
http://coleweb.dc.fi.udc.es/

© Springer-Verlag

Abstract. In this paper we study the performance of linguistically-motivated conflation techniques for Information Retrieval in Spanish. In particular, we have studied the application of productive derivational morphology for single word term conflation and the extraction of syntactic dependency pairs for multi-word term conflation. These techniques have been tested on several search engines implementing different indexing models. The aim of this study is to find the strong and weak points of each technique in order to develop heuristics for automatic query expansion.

1 Introduction

In Information Retrieval (IR) systems, documents are represented through a set of index terms or keywords. For such a purpose, documents are conflated by means of *text operations* [1, 6], which reduce their linguistic variety by grouping together textual occurrences referring to similar or identical concepts. However, most classical IR techniques for such tasks (such as the elimination of *stopwords*, too frequent words or words without seeming significance, or the use of *stemming*, which reduces distinct words to their supposed grammatical root) lack solid linguistic grounding. Even text operations with an apparent linguistic basis (e.g. stemming) which obtain very good results for English, perform badly when applied to languages with a very rich lexis and morphology, as in the case of Spanish. For these languages, we must face such tasks by employing Natural Language Processing (NLP) techniques, which redounds in a greater complexity and a higher computational cost.

2 NLP Techniques for Term Indexing

One of the main problems of natural language processing in Spanish is the lack of available resources: large tagged corpora, treebanks and advanced lexicons are not freely available. In this context, we propose to extend classical IR techniques in two ways: firstly, at word level, using morphological families; and secondly, at phrase level, using groups of related words with regard to their syntactic structure.

2.1 Morphological Families

Single word term conflation is usually accomplished in English through a *stemmer* [8], a simple tool from a linguistic point of view, with a low computational cost. The results obtained are satisfactory enough since the inflectional morphology of English is very simple. The situation for Spanish is completely different, because inflectional modifications exist at multiple levels with many irregularities [10]. The case of generative morphology is similarly very rich and complex in Spanish [7].

Using a lemmatizer we can solve the problems derived from inflection in Spanish. As a second step, we have developed a new approach based on morphological families [9]. We define a *morphological family* as a set of words obtained from the same morphological root through derivation mechanisms. It is expected that a basic semantic relationship will remain between the words of a given family.

For single word term conflation via morphological families, we first obtain the part of speech and the lemmas of the text to be indexed. Next, we replace each of the lemmas obtained by the representative of its morphological family. In this way, we are covering relations between terms of the type process-result, e.g. *producción* (production) / *producto* (product), process-agent, e.g. *manipulación* (manipulation) / *manipulador* (manipulator), and similar ones. These relations remain in the index because related terms are conflated to the same index term.

2.2 Syntactic and Morpho-Syntactic Variants

A *multi-word term* is a term containing two or more content words (nouns, adjectives and verbs). There exist several techniques to obtain them. The first one is *text simplification*: in a first step, we make a single word stemming, after which stopwords are deleted; in the final step, terms are extracted and conflated employing pattern matching [2] or statistical criteria [3]. As we can see, most operations lack solid linguistic grounding, which often results in incorrect conflations. Nevertheless, this is the easiest and least costly method.

At the other extreme, we find the *morpho-syntactic analysis* of the text by using a parser that produces syntactic trees which denote dependency relations between involved words. This way, structures with similar dependency relations are conflated in the same way.

At the mid point, we have *syntactic pattern matching*, which is based on the hypothesis that the most informative parts of the texts correspond to specific syntactic patterns [5].

We take an approach which conjugates these two last solutions, based on indexing noun syntagmas and their *syntactic and morpho-syntactic variants* [4].

A syntactic or morpho-syntactic variant of a multi-word term is a textual utterance that can be substituted for the original term in a task of information access:

Syntactic variants result from the inflection of individual words and from modifying the syntactic structure of the original term, e.g. *chico gordo* (fat boy) → *chicos gordos y altos* (fat and tall boys).

Morpho-syntactic variants differ from syntactic variants in that at least one of the content words of the original term is transformed into another word derived from the same morphological stem, e.g. *medición del contenido* (measurement of the content) → *medir el contenido* (to measure the content).

From a morphological point of view, syntactic variants refer to inflectional morphology, whereas morpho-syntactic variants also refer to derivational morphology. In the case of syntax, syntactic variants have a very restricted scope, a noun syntagma, whereas morpho-syntactic variants can span a whole sentence, including a verb and its complements, e.g. *comida de perros* (dog food) → *los perros comen carne* (dogs feed on meat). However, both variants can be obtained through transformations from noun syntagmas.

To extract such index terms we will use syntactic matching patterns obtained from the syntactic structure of the noun syntagmas and their variants. For such a task we take as our basis an approximate grammar for Spanish.

2.3 Syntactic and Morpho-Syntactic Variants as a Text Operation

The first task to be performed when indexing a text is to identify the index terms. Taking as our basis the syntactic trees corresponding to noun syntagmas and according to an approximate grammar for Spanish, we apply the mechanisms associated with syntactic and morpho-syntactic variants, obtaining their syntactic trees. Then, these trees are flattened into regular expressions formed by the part of speech labels of the words involved. Such matching patterns will be applied over the tagged text to be indexed, to identify the index terms. In this way, we are dealing with the problem from a surface processing approach at lexical level, leading to a considerable reduction of the running cost.

Once index terms have been identified, they must be conflated. This process consists of two phases. Firstly, we identify relations between pairs of content words inside the multi-word term, to conflate it into syntactic dependency-pairs. Secondly, single word term conflation mechanisms (lemmatization or morphological families) are applied to the words which form such pairs.

The relations we can find in a multi-word term correspond to three types:

1. *Modified-Modifier*, found in noun syntagmas. A dependency-pair is obtained for each combination of the head of the modifiers with the head of the modified terms. For example, *coches y motos rojas* is conflated into *(coche,rojo),(moto,rojo)* [1].

[1] *red cars and bikes* and *(car, red), (bike, red)*, respectively

2. *Subject-Verb*, relating the head of the subject and the verb.
3. *Verb-Complement*, relating the verb and the head of the complement.

In the case of syntactic variants, the dependencies of the original term always remain in the variant. In the case of morpho-syntactic variants we cannot guarantee the presence of the original term dependencies unless morphological families are applied. For example, given the term *recorte de gastos* (spending cutback) and its morpho-syntactic variant *recortar gastos* (to cut back spending), using lemmatization we obtain the following two different dependency pairs *(recorte, gasto)* and *(recortar, gasto)*, respectively. Whereas, using morphological families and supposing that the representatives are *recorte* (cutback) and *gastar* (to spend), we obtain the same dependency pair *(recorte, gastar)* for both the original term and its variant. Therefore, the degree of conflation obtained when using morphological families is higher than when using lemmatization.

3 Evaluation

The techniques proposed in this paper are independent of the indexing engine we choose to use because documents are preprocessed before being treated. We have performed experiments using the following engines: Altavista SDK[2], SMART[3] (based on a vector model) and SWISH-E[4] (based on a boolean model). For all engines, we have tested five different conflation techniques:

1. Elimination of stopwords using the list provided by SMART (*pln*).
2. Lemmatization of content words (*lem*).
3. Morphological families of content words (*fam*).
4. Syntactic dependency-pairs with lemmatization (*FNL*).
5. Syntactic dependency-pairs with morphological families (*FNF*).

We have tested the five proposed approaches on a corpus of 21,899 documents of a journalistic nature (national, international, economy, culture, ...) covering the year 2000. The average length of the documents is 447 words. We have considered a set of 14 natural language queries with an average length of 7.85 words per query, 4.36 of which were content words.

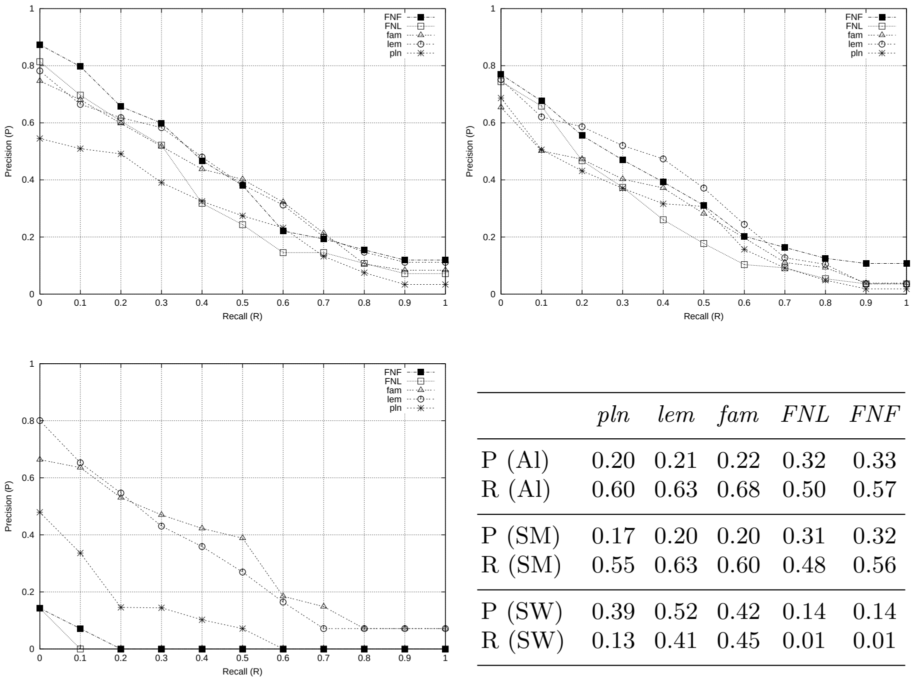
3.1 Altavista SDK

Average precision and recall are shown at the bottom-right of Fig. 1. Single word term conflation techniques, *fam* and *lem*, has led to a remarkable increase in precision and recall with respect to *pln.* The increase in precision is even higher for multi-word term conflation techniques *FNL* and *FNF*. Moreover, the technique *FNF* attains good recall, which implies that it can significantly increase the precision without seriously affecting the recall.

[2] `http://solutions.altavista.com/`

[3] `ftp://ftp.cs.cornell.edu/pub/smart/`

[4] `http://sunsite.berkeley.edu/SWISH-E/`



	pln	*lem*	*fam*	*FNL*	*FNF*
P (Al)	0.20	0.21	0.22	0.32	0.33
R (Al)	0.60	0.63	0.68	0.50	0.57
P (SM)	0.17	0.20	0.20	0.31	0.32
R (SM)	0.55	0.63	0.60	0.48	0.56
P (SW)	0.39	0.52	0.42	0.14	0.14
R (SW)	0.13	0.41	0.45	0.01	0.01

Fig. 1. *Global results: Altavista (top-left), SMART (top-right), SWISH-E (bottom-left) and average precision and recall (bottom-right)*

With respect to the evolution of precision vs. recall, Fig.1 confirms the technique *pln* as being the worst one. The best behaviour corresponds to the technique *FNF*, except for the segment of recall between 0.5 to 0.7, where single word term conflation techniques, *lem* and *fam*, are slightly better. For low recall rates (≤ 0.3) *FNF* is clearly the best one, whilst the other conflation techniques show a similar behaviour. For the segment of recall between 0.3 to 0.5, single word term conflation techniques are closer to *FNF*, whereas *FNL* is closer to *pln*. For recall rates greater than 0.7, conflation techniques using lemmatization tend to converge, as do the techniques using morphological families.

3.2 SMART

The average recall and precision are similar to those obtained with Altavista SDK, except for the case of single word term conflation via morphological families, *fam*, which does not appear to improve the global behaviour of the system in relation to lemmatization. On the contrary, its efficiency is somewhat reduced, in contrast with the results for Altavista. Nevertheless, the use of such families together with the use of multi-word terms gives a remarkable increase of recall, as in the case of Altavista, with regard to the use of lemmatization with complex terms. In fact, *FNF* improves the recall of *pln.* We can also notice that there is a greater homogeneity in the behaviour of all methods in the case of recall.

With respect to the evolution of precision vs. recall, we can observe in Fig. 1 that the greater complexity of the vectorial model tends to reduce the differences between all techniques. We can observe some noticeable differences between the behaviour of conflating techniques in SMART and Altavista. Firstly, in SMART the behaviour of *fam* technique is clearly worse than the behaviour of *lem* technique. This supports the results obtained for average precision and recall. Another difference we can observe is that in SMART the *FNF* technique only obtains better results than the rest of methods for low and high recalls (≤ 20, $50 \geq$), while for the rest of the interval the best method is clearly *lem.* On the other hand, we also find some similarities, such as the fact that *pln* and *FNL* have the worst behaviour in comparison with the other methods.

3.3 SWISH-E

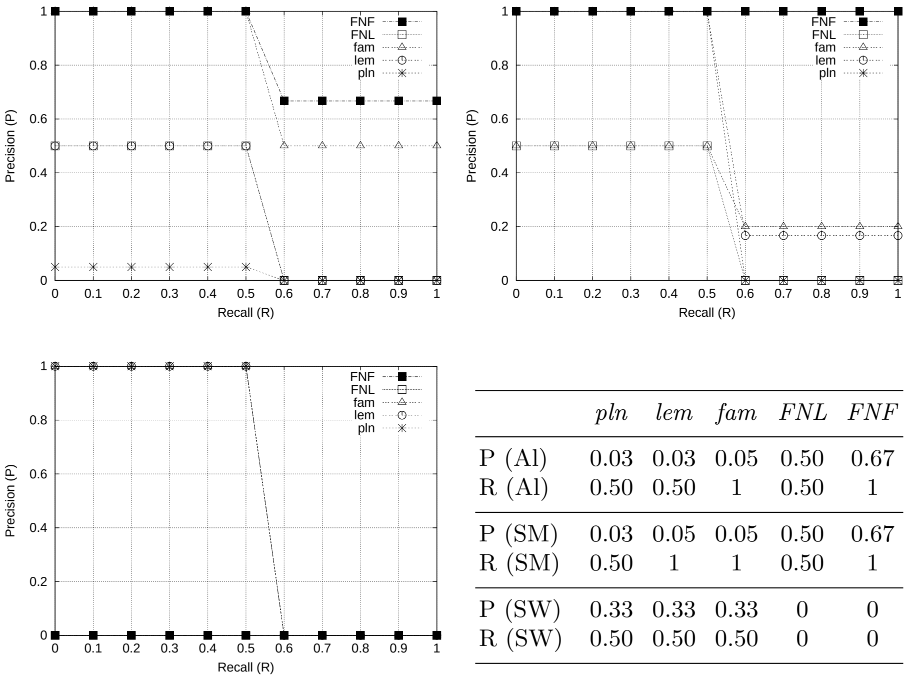
The first conclusion we can reach is that the use of multi-word term methods in combination with the boolean model is completely inadequate due to boolean engines require all terms involved in a query to match index terms in a given document, a rare situation when dealing with syntactic dependencies. The use of plain text with a boolean model is also completely inadequate because this model is more sensitive to inflectional variations than the previous engines. When using lemmatization to conflate the text we reach a noticeably higher level of recall, with very high precision. The employment of morphological families for single word term conflation obtains a higher level of recall, but the level of precision is lower than the level of precision attained with *lem.* This is due to the noise introduced by inaccurate families. However, it is also interesting to remark that the precision reached by *pln*, *lem* and *fam* is the highest reached for all the test suite, but at the cost of reducing recall.

3.4 Behaviour for Particular Queries

The behaviour of the different techniques varies according to the characteristics of each particular query. We will try to illustrate this fact with some practical examples obtained during the test process. This study is a first step towards the development of heuristics for automatic query expansion.

The first example we will work with is the query *"experimentos sobre la clonación de monos"* (experiments on the cloning of monkeys), trying to illustrate how multi-word terms can discriminate very specific information. For this query only two relevant documents were found in the corpus. Nevertheless, since cloning is a very popular topic nowadays, there are a lot of related but non-relevant documents which introduce a lot of noise.

We center this discussion on the results obtained by multi-word term conflation techniques. As we can see in the graph of precision vs. recall for Altavista and SMART of Fig. 2 the evolution for techniques based on families (*fam* and *FNF*) is similar, and the same occurs with techniques using lemmatization (*lem*

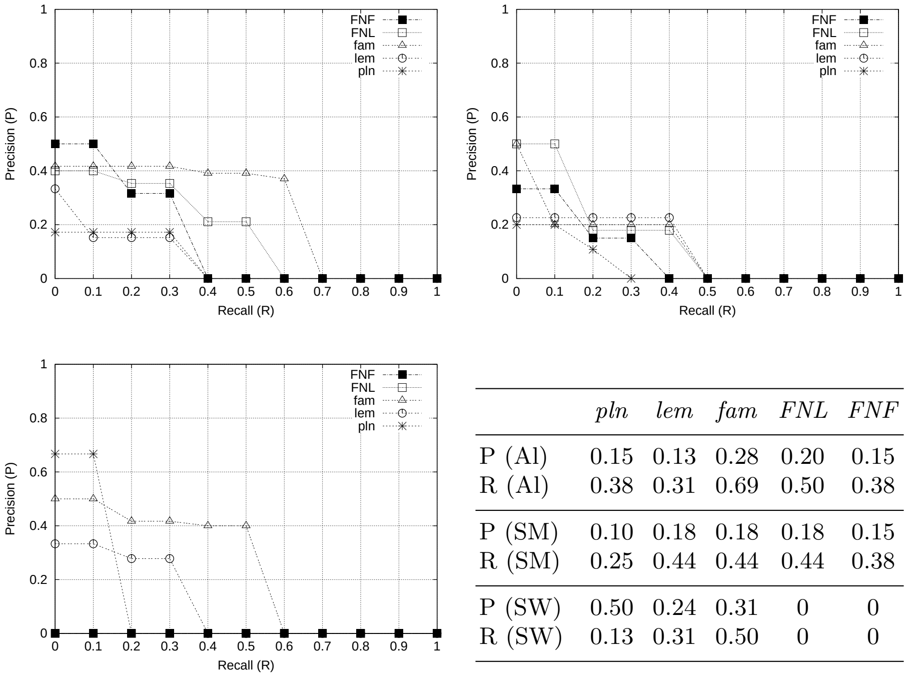


	pln	*lem*	*fam*	*FNL*	*FNF*
P (Al)	0.03	0.03	0.05	0.50	0.67
R (Al)	0.50	0.50	1	0.50	1
P (SM)	0.03	0.05	0.05	0.50	0.67
R (SM)	0.50	1	1	0.50	1
P (SW)	0.33	0.33	0.33	0	0
R (SW)	0.50	0.50	0.50	0	0

Fig. 2. *"experimentos sobre la clonación de monos"*: results for Altavista (top-left), SMART (top-right), SWISH-E (bottom-left) and average precision and recall

and *FNL*). Nevertheless these last two techniques obtain a lesser degree of recall and precision. However, it is in the table of average precision and recall where we can find very important differences. We can see that the average precision reached by using multi-word term conflation techniques is significatively higher than the average precision obtained by using single word term techniques. Morover, the levels of recall are maintained. This means that the set of documents returned is small but precise, because multi-word term techniques have been able to discriminate the relevant documents adequately without losing recall.

As a second example, we consider the query *"negociaciones del PP con el PSOE sobre el pacto antiterrorista"* (negotiations between PP and PSOE about the pact against terrorism) to illustrate a case where single word term conflation techniques achieve a better performance than multi-word techniques. As we can observe in figure 3, there are some similarities in the precision vs. recall graphs for the three indexing engines. In particular, the *fam* technique shows the best behaviour, even better than multi-word techniques. The *lem* technique performs worse than *fam* but better than *pln* for all search engines. These results are due to the fact that lemmatization solves variations caused by inflectional morphology. In addition, morphological families solve variations caused by derivational morphology, retrieving more documents, most of which are relevant because this



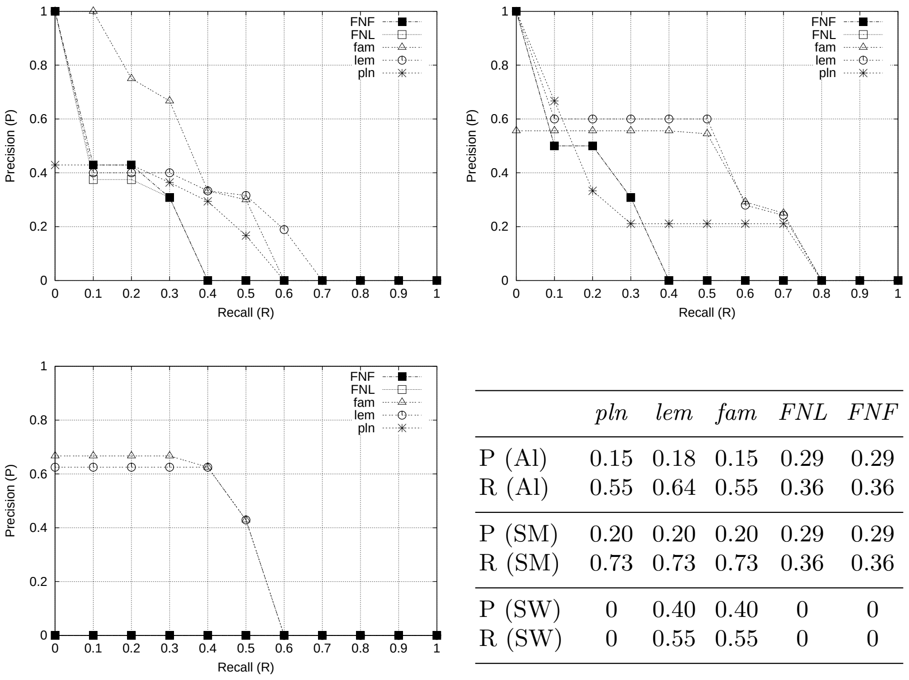
	pln	*lem*	*fam*	*FNL*	*FNF*
P (Al)	0.15	0.13	0.28	0.20	0.15
R (Al)	0.38	0.31	0.69	0.50	0.38
P (SM)	0.10	0.18	0.18	0.18	0.15
R (SM)	0.25	0.44	0.44	0.44	0.38
P (SW)	0.50	0.24	0.31	0	0
R (SW)	0.13	0.31	0.50	0	0

Fig. 3. *"negociaciones del PP con el PSOE sobre el pacto antiterrorista"*: Altavista (top-left), SMART (top-right), SWISH-E (bottom-left) and average precision and recall

query involves words with derivatives which frequently appear in the texts referring to the topic of the query, such as *negociación* (negotiation), *negociar* (to negotiate) and *negociador* (negotiator) or *pacto* (pact) and *pactar* (to agree on).

The third example we consider corresponds to the query *"el PSOE reclama un debate entre Aznar y Almunia"* (PSOE demands a debate between Aznar and Almunia). This query refers to the constant demand by the PSOE party for a TV debate between the two main candidates in the general elections in the year 2000. We must take into account that words like *PSOE*, *Almunia*, *Aznar* and *debate* appear in several hundreds of non-relevant documents about political issues, introducing a lot of noise. Finally, we have found that only 11 documents were relevant to this query. With this example we try to illustrate the situations where a boolean model beats other indexing models, as is shown in Fig. 4.

Comparing the precision vs. recall graph for SWISH-E with respect to the graphs for Altavista and SMART, we observe that precision in these two last models is lower or similar to precision in SWISH-E for all levels of recall, except for the interval ≤ 0.1. However, the main difference arises in the measures of average precision and recall. Recall in SWISH-E reaches 55%, in contrast to 64% and 73% reached by Altavista and SMART, respectively. Nevertheless, precision in SWISH-E reaches 40%, in contrast to a maximum of 29% reached by multi-



	pln	*lem*	*fam*	*FNL*	*FNF*
P (Al)	0.15	0.18	0.15	0.29	0.29
R (Al)	0.55	0.64	0.55	0.36	0.36
P (SM)	0.20	0.20	0.20	0.29	0.29
R (SM)	0.73	0.73	0.73	0.36	0.36
P (SW)	0	0.40	0.40	0	0
R (SW)	0	0.55	0.55	0	0

Fig. 4. *"el PSOE reclama un debate entre Aznar y Almunia"*: results for Altavista (top-left), SMART (top-right), SWISH-E (bottom-left) and average precision and recall

word term conflation techniques in the other engines. This increase in precision is due to the high level of discrimination achieved by the boolean model between relevant and non-relevant documents when the query is very similar to the way the information is expressed in the documents, i.e. there exists little variation in the way the concepts involved in the query are expressed.

4 Conclusions

We have shown how linguistically-motivated indexing can improve the performance of information retrieval systems working on languages with a rich lexis and morphology, such as Spanish. In particular, two text operations have been applied to effectively reduce the linguistic variety of documents: productive derivational morphology for single word term conflation and extraction of syntactic dependency-pairs for multi-word term conflation. These techniques require a minimum of linguistic resources, which make them adequate for processing European minority languages. The increase of computational cost is also minimal due to the fact that they are based on finite state technology, which makes them useful for practical systems.

These techniques have been tested on a testsuite of journalistic documents using different search engines. We have found that:

- Indexing of plain text (*pln*) is the worst option, independently of the indexing model used.
- Morphological families show good recall and precision when they do not introduce noise.
- Multi-word term conflation techniques (*FNL*,*FNF*) do not work properly in combination with the boolean model.
- Multi-word term conflation significantly increases the precision in non-boolean models, and when combined with morphological families (*FNF*) it also shows a good level of recall.
- Lemmatization (*lem*) is not the best technique but it is a good balance between all the techniques considered.

In consequence, we can propose an automatic heuristic consisting in the use of morphological families (*fam*) with a boolean model when the words involved in the query have variants with a high frequency of appearance in the corpus of documents. Depending on user need, we can also propose the following heuristics for interaction between the information retrieval system and the user:

- When the user is searching nearly literal utterances, the best option is to employ lemmatized text (*lem*) with a boolean search engine.
- If the user requires high precision, even at the expense of reducing recall, dependency-pairs with families (*FNF*) is the most accurate approach.
- When the user wishes to increase recall, for example if the other techniques return few documents, he may use morphological families (*fam*).

Acknowledgments

This research has been partially supported by Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (Grant TIC2000-0370-C02-01), FEDER of EU (Grant 1FD97-0047-C04-02) and Xunta de Galicia (Grants PGIDT99XI10502B and PGIDT01PXI10506PN).

References

1. R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern information retrieval.* Addison-Wesley, Harlow, England.
2. M. Dillon and A.S. Gray. 1983. FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2):99–108.
3. J.L. Fagan. 1987. Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *Proceedings of ACM SIGIR'87*, pages 91–101.

4. C. Jacquemin and E. Tzoukerman. 1999. NLP for term variant extraction: A synergy of morphology, lexicon and syntax. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25–74. Kluwer Academic, Boston.
5. J.S. Justeson and S.M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.
6. G. Kowalski. 1997. *Information retrieval systems: theory and implementation.* Kluwer Academic, Boston.
7. M. F. Lang. 1990. *Spanish Word Formation: Productive Derivational Morphology in the Modern Lexis.* Croom Helm. Routledge, London and New York.
8. M. Lennon, D.S. Pierce, and P. Willett. 1981. An evaluation of some conflation algorithms. *Journal of Information Science*, 3:177–183.
9. J. Vilares, D. Cabrero, and M. A. Alonso. 2001. Applying Productive Derivational Morphology to Term Indexing of Spanish Texts. In Proc. 2nd International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2001, Mexico City, Mexico; February 18–24. To be published in *Lecture Notes in Artificial Intelligence.*
10. M. Vilares, J. Graña, and P. Alvariño. 1997. Finite-state morphology and formal verification. In A. Kornai, editor, *Extended Finite State Models of Language*, pages 37–47. Cambridge University Press.