Using syntactic dependency-pairs conflation to improve retrieval performance in Spanish*

Jesús Vilares, Fco. Mario Barcala, and Miguel A. Alonso

Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 La Coruña, Spain
{jvilares,barcala}@mail2.udc.es alonso@udc.es
http://coleweb.dc.fi.udc.es/

Abstract. This article presents two new approaches for term indexing which are particularly appropriate for languages with a rich lexis and morphology, such as Spanish, and need few resources to be applied. At word level, productive derivational morphology is used to conflate semantically related words. At sentence level, an approximate grammar is used to conflate syntactic and morphosyntactic variants of a given multi-word term into a common base form. Experimental results show remarkable improvements with regard to classical indexing methods.

1 Introduction

For Information Retrieval (IR) tasks, documents are frequently represented through a set of index terms or representative keywords. This can be accomplished through operations such as the elimination of *stopwords* (too frequent words or words with no apparent significance) or the use of *stemming* (which reduces distinct words to their supposed grammatical root). These operations are called *text operations*, providing a *logical view* of the processed document.

In effect, current IR systems conflate the documents before indexing to decrease their linguistic variety by grouping together textual occurrences referring to similar or identical concepts by exploiting graphical similarities, thesaurus, etc. [1, 7]. However, most classical IR techniques for such tasks lack solid linguistic grounding. Even operations with an apparent linguistic basis (e.g. stemming) which obtain good results for English, perform badly when applied to languages with a very rich lexis and morphology, such as Spanish. For these languages, we must employ more and better linguistic resources with Natural Language Processing (NLP) techniques, all of which involves a greater complexity and a higher computational cost. At this point, we must face one of the main problems of NLP in Spanish, which 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 to avoid such obstacles.

* 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).

2 Single word term conflation

In English, single word term conflation can be accomplished through a *stemmer* [9], 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 being very simple. The situation for Spanish is completely different, because inflectional modifications exist at multiple levels[1] with many irregularities. Therefore, we must apply NLP techniques, thus increasing the complexity and the computational cost of the system. As a first step, we have employed a lemmatizer to obtain the lemma of each word, thereby solving the problems derived from inflection in Spanish. As a second step, we have developed a new approach based on morphological families.

2.1 Morphological families as a text operation

Spanish has a great productivity and flexibility in its word formation mechanisms by using a rich and complex productive morphology, preferring derivation to other mechanisms [2]. 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[2]. Regular word formation patterns in Spanish can be obtained through the 'rules of word formation' [8] defined by generative phonology and transformational-generative grammars. Though this paradigm is not complete, it can be used to implement an automatic system for generation of morphological families with an acceptable degree of completeness and correction [10].

In order to use morphological families for document conflation, the first step is to 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 using the same index term to represent all words belonging to the same morphological family; therefore, semantic relations that exist between these words remain in the index because related terms are conflated to the same index term.

We have compared the accuracy of lemmatization and morphological families as text operations with respect to the classical technique of stemming. We have studied the behaviour of different stemmers specifically designed for Spanish, and the best results we obtained were for the stemmer used by the open source search engine Muscat[3], based on Porter's algorithm [1]. However, such results were poor. The employment of a lemmatizer allowed us to reach an approximate accuracy of 96%, whereas the Muscat stemmer only reached 37% overall. Furthermore, the behaviour of a lemmatizer is uniform for all grammar categories,

[1] Gender and number for nouns and adjectives, and person, mood, time and tense for verbs.

[2] Relations of the type process-result, e.g. *producción* (production) / *producto* (product), process-agent, e.g. *manipulación* (manipulation) / *manipulador* (manipulator), etc.

[3] `http://open.muscat.com`

whereas stemmers obtain an accuracy of 46% for nouns, 36% for adjectives and 0% for verbs[4]. A noticeable extra advantage of lemmatizers in relation to stemmers is their capability to disambiguate using word context. Moreover, comparing stemmers with respect to morphological families, we find that the Muscat stemmer is able to identify 27% of the families, 95% of which are families formed by only one lemma, 3% by two lemmas, and less than 2% by three lemmas.

With regard to computational cost, morphological families and their representatives are computed *a priori*, so they do not affect the final indexing and querying cost. The running cost of a stemmer is linear in relation to the length of the word. The running cost of a lemmatizer-disambiguator is only slightly greater: linear in relation to the length of the word and cubic in relation to the size of the tagset, which is a constant. As will be detailed in Sect. 3.3, our system only needs to know the grammatical category of the word, so the tagset will be very small. Therefore, the increase in cost becomes negligible.

3 Multi-word term conflation

A *multi-word term* is a term containing two or more content words (nouns, verbs and adjectives) [5]. Several techniques are described in the literature to obtain them. One of the most frequently used is *text simplification* [5]: as a first step, we make a single word stemming, after which stopwords are deleted; in the final step, terms are extracted and conflated by means of pattern matching [3], statistical criteria [4], etc. As we can see, most operations lack solid linguistic grounding[6], 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, which uses a parser that produces syntactic trees which denote dependency relations between involved words. As a result, 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 [6]. In this article we take an approach that combines these two last solutions, trying to obtain the concepts of a text by means of the syntactic relations that exist between the terms of the document. These syntactic relations will be identified through *syntactic patterns* of noun syntagmas and their *syntactic and morpho-syntactic variants*.

A syntactic or morpho-syntactic variant of a multi-word term is a textual utterance in which:

– Syntactic variants result from the inflection of individual words and from modifying the syntactic structure of the original term. E.g. *chicos gordos y altos* (fat and tall boys) is a variant of *chico gordo* (fat boy).

[4] This is due to the complexity of the verbal paradigm in Spanish, which is not treated in depth by any stemmer.

[5] E.g. *el perro grande del vecino* (the neighbour's big dog)

[6] For example, stopwords such as determiners and prepositions are key components of the syntactic structure.

– 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. *medir el contenido* (to measure the content) is a variant of *medición del contenido* (measurement of the content).
– The original term can substitute the variant in a task of information access.

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, i.e. a noun syntagma, whereas morpho-syntactic variants can span a whole sentence, including a verb and its complements[7]. Next, we will study the mechanisms involved in obtaining syntactic and morpho-syntactic variants.

3.1 Syntactic variants

In Spanish, syntactic variants of a multi-word term may involve variations in the inflection of its words, and syntactic alterations of the kind:

– *Coordination:* this consists of employing coordinating constructions (copulative or disjunctive) with the modifier or with the modified term. For example, *coches rojos* (red cars) and *motos rojas* (red bikes) combine into *coches y motos rojos* (red cars and bikes), which can be considered as a variant of any of the combined terms.
– *Substitution:* it consists of employing modifiers to make a term more specific. For example, *caída en las ventas* (sales drop) can be transformed into *caída anormal en las ventas* (unusual sales drop) by adding the adjective *anormal*.
– *Synapsy:* whereas the preceding constructions are binary, this is a unary construction which corresponds to a change of preposition or the addition or removal of a determiner. For example, we can obtain *abono para plantas* (fertilizer for plants) from *abono para las plantas* (fertilizer for the plants).
– *Permutation:* this refers to the permutation of words around a pivot element, for example *saco viejo* (old bag) and *viejo saco* (old bag).

3.2 Morpho-syntactic variants

According to the nature of the morphological transformations applied to the content words of the terms, we can classify morpho-syntactic variants into:

– *Iso-categorial:* the morphological derivation process does not change the category of the word, but only transforms one noun syntagma into another. There are two possibilities:
  1. *Noun-to-Noun:* they cover relations of the type process-result —*producción artesanal* (craft production) / *producto artesanal* (craft product)— and process-agent —*manipulación de las masas* (manipulation of the masses) / *manipulador de las masas* (manipulator of the masses)—.

[7] Let us consider *comida de perros* (dog food) and *los perros comen* (dogs feed on).

2. *Adjective-to-Adjective:* covering relations of the type agent-result — *compuesto ionizador* (ionizer compound) / *compuesto ionizado* (ionized compound)—.

– *Hetero-categorial:* morphological derivation does result in a change of the category of the word. They are not restricted to the frontier of a noun syntagma.
  1. *Noun-to-Verb:* these variations involve semantic changes of the type process-result, e.g. *recortar gastos* (to cut back spending) / *recorte de gastos* (spending cutback).
  2. *Noun-to-Adjective:* in a noun syntagma the noun can be modified by adjectival constructions or equivalent prepositional ones, e.g. *cambio del clima* (change of the climate) / *cambio climático* (climatic change).

3.3 Term extraction and conflation

In information systems, many of the queries can be formulated as noun syntagmas of diverse complexity. Thus, we will take noun syntagmas as base terms from which we will obtain, through the corresponding mechanisms, their syntactic and morpho-syntactic variants, not necessarily noun syntagmas. All these multi-word terms, either the original noun syntagmas or their variants, can be used as index terms.

In Spanish, the basic structures for noun syntagmas are four: *Adj-Noun*, *Noun-Adj*, *Noun-Prep-Noun* and *Noun-Prep-Det-Noun*. So, we are interested in identifying such noun syntagmas and their variants for indexing.

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:

$$
\begin{array}{lll}
S & \rightarrow NP\ V\ W^{?}\ (NP|PP)^{*} & (1) \\
NP & \rightarrow D^{?}\ AP^{*}\ N\ (AP|PP)^{*} & (2) \\
AP & \rightarrow W^{?}\ A & (3) \\
PP & \rightarrow P\ NP & (4)
\end{array}
$$

where the symbols D, A, N, W, V and P are the part of speech labels that denote determiners, adjectives, nouns, adverbs, verbs and prepositions, respectively[8]. The motivation of these rules is:

(1) shows a sentence structure of the kind *Subject-Verb-Complement*.
(2) defines a noun syntagma as a noun modified by adjectives and/or prepositional syntagmas.
(3) lets adjectives be modified by adverbs.
(4) shows a prepositional syntagma formed by a preposition and a noun syntagma.

[8] Coordinating conjunctions (C) and punctuation marks (Q) will be also used later to obtain variants.

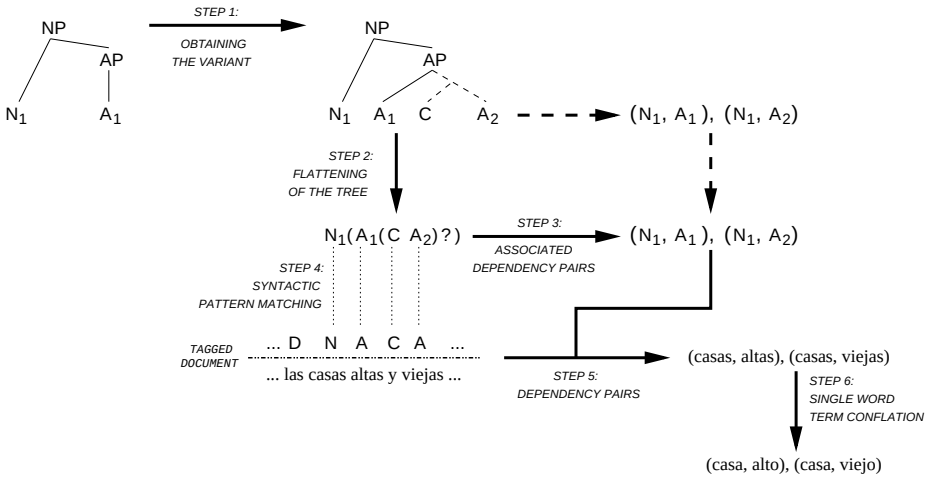

Fig. 1. Example of multi-word term conflation via dependency pairs

Other authors, such as [5], take a static approach based on the use of previous existing terminological databases, which are incorporated into a lexicalized parser. Since this kind of resources is very difficult to obtain for Spanish, we opt for a dynamic approach in which terms are dynamically identified during the indexing process without any deep syntactic processing of the document, only a surface process, this approach having no terminological reference at all. In this way, the increase of computational cost and the number of extra linguistic resources employed by the system are minimal, key questions for being employed in real-world applications.

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 the approximate grammar we have previously shown, we manually apply the mechanisms described in Sects. 3.1 and 3.2. As a result, we obtain the syntactic trees corresponding to syntactic and morpho-syntactic variants of such noun syntagmas. This set of trees that we have obtained for multi-word terms (noun syntagmas and their variants) can be classified into four main groups: *noun modified by adjectives*, *noun modified by prepositional syntagmas*, *verb-complement* and *subject-verb*.

However, in our approach, these trees are not directly applicable to term extraction. First, they are flattened into regular expressions using the part of speech labels of the tokens involved. Let us take the example shown in Fig. 1:

1. We start with a noun syntagma whose syntactic structure is shown in the left tree, with the head noun N_1 modified by an adjectival syntagma.
2. We obtain one of its variants through the incorporation of a coordination into the adjectival syntagma (*step 1*).
3. The syntactic tree of the obtained variant is flattened to obtain the pattern which will be applied to the tagged text (*step 2*).

	source	*pln*	*lem*	*fam*	*FNL*	*FNF*
Total	9,780,513	4,526,058	4,625,579	4,625,579	2,666,190	2,666,190
Unique	154,419	154,071	111,982	105,187	1,210,182	1,036,005

Table 1. Statistics of the composition of the test corpus

Once index terms have been identified through syntactic matching patterns, they must be conflated. This process consists of two phases. Firstly, we identify syntactic dependencies between pairs of content words inside the syntactic tree of the multi-word term (*syntactic-dependency pairs*); such pairs are now associated with the matching pattern which corresponds with that tree. Secondly, single word term conflation mechanisms (lemmatization or morphological families) are applied to the words which form such pairs; the resultant pairs are the terms to be indexed.

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

1. *Modified-Modifier:* these kinds of relation are 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:
   chicos feos y altos → (chico, feo),(chico, alto)
   (ugly and tall boys → (ugly, boy), (tall, boy))
2. *Subject-Verb:* the main dependency is the one relating the head of the subject and the verb. For example:
   los perros comen carne → (perro, comer)
   (dogs feed on meat→ (dog, to feed on))
3. *Verb-Complement:* the main dependency is the one relating the verb and the head noun of the complement. For example:
   recortar gastos → (recortar, gasto)
   (to cut back spending → (to cut back, spending))

In Fig. 1, the dependency pairs associated with the variant are obtained in *step 3*.

In the case of syntactic variants, the dependencies of the original multi-word term always remain in the variant. Nevertheless, in the case of morpho-syntactic variants, this only happens when morphological families are applied to conflate the single word terms of the pair. 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 pairs *(recorte, gasto)* and *(recortar, gasto)*, respectively. Nevertheless, using morphological families we obtain the same dependency pair *(recorte, gastar)* for both the original term and its morpho-syntactic variant[9]. Therefore, the degree of conflation we obtain using morphological families is higher than using lemmatization.

[9] In this example we have supposed that *recorte* is the representative of the family of *recorte* and *recortar*, whereas *gastar* is the representative of the family of *gasto* and *gastar*

To end our explanation we can also see in Fig. 1 an example of the conflation process of the term *casas altas y viejas* (tall and old houses) using the structures previously obtained. In *step 4* tagged text is matched with the pattern, to obtain in *step 5* its associated dependency pairs. Finally, in *step 6* single terms forming each pair are conflated, obtaining the actual pairs to be indexed.

4 Evaluation of the system

The techniques proposed in this article are independent of the indexing engine we choose to use. This is because we first conflate each document to obtain its index terms; then, the engine receives the conflated version of the document as input. So, any standard text indexing engine may be employed, which is a great advantage. Nevertheless, each engine will behave according to its own characteristics [10].

For evaluating the system, five indexing methods have been tested:

pln: plain text eliminating stopwords.
lem: single word term conflation via lemmatization.
fam: single word term conflation via morphological families.
FNL: multi-word term conflation via syntactic dependency-pairs and lemmatization.
FNF: multi-word term conflation via syntactic dependency-pairs and morphological families.

The corpus used for evaluation is formed by 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.

Table 1 shows the statistics of the terms that compose this corpus. The first and second row show the total number of terms and unique terms obtained for the indexed documents, respectively, either for the source text and for the different conflated texts. As we can observe in the upper row, single word term conflation techniques attain a reduction of more than 50% in the number of terms to index whereas multi-word term conflation techniques attain a reduction of nearly 75%. With respect to the number of different terms of the indexes, shown in the lower row, the reduction provided by the elimination of stopwords is negligible, whereas lemmatization and morphological families provide a reduction of 27% and 32%, respectively, with the consequent saving of space and reduction of accessing time to the indexes. Moreover, multi-word term conflation techniques significantly increase the number of index terms since they are complex terms which express syntactic relations. However, we must point out that the use of morphological families to construct such complex terms reduces the number of index terms with respect to the use of lemmatization by 14%, whereas their employment for single word term conflation only attained a relative extra reduction of 6%.

[10] Indexing model, ranking algorithm, etc.

	pln	*lem*	*fam*	*FNL*	*FNF*
Average precision	0.1714	0.2018	0.1982	0.3050	0.3215
Average recall	0.5515	0.6316	0.6028	0.4788	0.5615

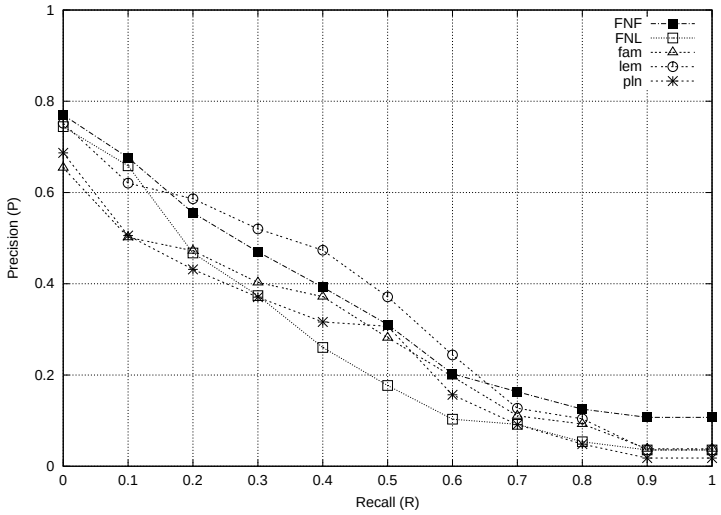

Fig. 2. Average precision and recall and precision vs. recall graph

The results we show in this section have been obtained for the vector-based search engine SMART[11]. In Fig. 2 you can find the results obtained for average recall and precision.[12] We can observe that the application of techniques for single word term conflation, *fam* and *lem*, has led to a remarkable increase in recall whereas the techniques for multi-word term conflation, *FNL* and *FNF*, has led to a remarkable increase in precision. It should be noticed that the isolated employment of morphological families (*fam*) does not always guarantees improvements with respect to lemmatization (*lem*). However, its employment together with multi-word terms (*FNF*) attains a noticeable increase in recall with respect to lemmatization (*FNL*).

With respect to the evolution of precision vs. recall, Fig. 2 confirms the technique *pln* as being the worst one, whereas the best behavior corresponds to *lem* and *FNF*. For low and high recall rates (≤ 0.2, ≥ 0.7) *FNF* is clearly the best one, whereas for the rest of the interval *lem* does better.

5 Conclusions

In this article we have shown how linguistically-motivated indexing can improve the performance of Information Retrieval (IR) systems working on languages

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

[12] Results for individual queries depend on the characteristics of each query [11].

with a rich lexis and morphology, such as Spanish. In particular, two new text operations to effectively reduce the linguistic variety of documents have been applied: productive derivational morphology for single word term conflation and syntactic dependency-pairs obtained from approximate grammars for multi-word term conflation.

Unlike other related approaches based on parsing and large terminological databases, which gives them a static nature, our approach is dynamic since index terms are identified in running time. It also requires a minimum of linguistic resources, which makes it appropriate for processing European minority languages. As it is a lexical approach, the increase of computational cost is also minimum due to the fact that it is based on finite state technology, allowing its practical application in real systems.

Experimental results allow us to conclude that the isolated employment of morphological families does not always guarantees improvements with respect to lemmatization, but their use together with multi-word terms substantially increases precision whilst maintaining a very acceptable level of recall.

References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*. Addison-Wesley, Harlow, England, 1999.
2. Elena Bajo Pérez. *La derivación nominal en español*. Cuadernos de lengua española. Arco Libros, Madrid, 1997.
3. M. Dillon and A.S. Gray. FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2):99–108, 1983.
4. J.L. Fagan. Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *Proceedings of ACM SIGIR'87*, pages 91–101, 1987.
5. Christian Jacquemin and Evelyne Tzoukerman. NLP for term variant extraction: A synergy of morphology, lexicon and syntax. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25–74. Kluwer Academic, Boston, 1999.
6. J.S. Justeson and S.M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.
7. Gerald Kowalski. *Information retrieval systems : theory and implementation*. Kluwer international series on information retrieval. Kluwer Academic, Boston, 1997.
8. Mervyn F. Lang. *Spanish Word Formation: Productive Derivational Morphology in the Modern Lexis*. Croom Helm. Routledge, London and New York, 1990.
9. M. Lennon, D.S. Pierce, and P. Willett. An evaluation of some conflation algorithms. *Journal of Information Science*, 3:177–183, 1981.
10. Jesús Vilares, David Cabrero, and Miguel A. Alonso. Applying productive derivational morphology to term indexing of spanish texts. In *Computational Linguistics and Intelligent Text Processing*, LNCS 2004, pages 336–348. Springer-Verlag, 2001.
11. Jesús Vilares, Manuel Vilares, and Miguel A. Alonso. Towards the development of heuristics for automatic query expansion. In H. C. Mayr, J. Lazansky, G. Quirchmayr, and P. Vogel, editors, *Database and Expert Systems Applications*, LNCS 2113, pages 887–896. Springer-Verlag, Berlin-Heidelberg-New York, 2001.