COLE experiments at CLEF 2002 Spanish monolingual track *

Jesús Vilares
Miguel A. Alonso
Departamento de Computación
Universidade da Coruña
Campus de Elviña s/n
15071 La Coruña (Spain)
jvilares@mail2.udc.es
alonso@udc.es

Francisco J. Ribadas
Manuel Vilares
Escuela Superior de Ingeniería Informática
Universidade de Vigo
Campus As Lagoas s/n
32004 Orense (Spain)
ribadas@ei.uvigo.es
vilares@ei.uvigo.es

Abstract

In this our first participation in CLEF, we have applied Natural Language Processing techniques for single word and multi-word term conflation. We have tested several approaches at different levels of text processing in our experiments: firstly, we have lemmatized the text to avoid inflectional variation; secondly, we have expanded the queries through synonyms according to a fixed threshold of similarity; and thirdly, we have tested a mixed approach based on the employment of productive derivational morphology to solve derivational variation and syntactic dependencies to deal with the syntactic content of the document.

1 Introduction

In Text Retrieval, since the information is encoded as text, the task of deciding whether a document is relevant or not to a given information need can be viewed as a Natural Language Processing (NLP) problem, in particular for languages with rich lexical, morphological and syntactical structures, such as Spanish. Moreover, during recent years the progress in the field of NLP has resulted in the development of a new generation of more efficient, robust and precise tools. These advances, together with the increasing power of new computers, allow us to apply such NLP systems in real IR environments.

Nevertheless, at this point, we must face one of the main problems of NLP in Spanish, the lack of available linguistic resources: large tagged corpora, treebanks and advanced lexicons are not available. Therefore, while waiting for the availability of such resources, the only solution is to look for simplicity, employing a minimum of linguistic resources.

In this paper we present a set of NLP tools designed for dealing with different levels of linguistic variation in Spanish: morphological, lexical and syntactical. The effectiveness of our solutions has been tested during this our first participation in the CLEF Spanish monolingual track.

This article is outlined as follows. Section 2 describes the techniques used for single word term conflation. Expansion of queries by means of synonyms is introduced in Section 3. Multi-word term conflation through syntactic dependencies is described in Section 4. Section 5 shows non-official results obtained for CLEF 2001 queries, whereas Section 6 shows the official results for CLEF 2002.

2 Conflation of words using inflectional and derivational morphology

Our proposal for single word term conflation is based on exploiting the lexical level in two phases: firstly, by lemmatizing the text to solve inflectional variation, and secondly, by employing morphological families to deal

*The research reported in this article has been supported in part by Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (TIC2000-0370-C02-01), Ministerio de Ciencia y Tecnología (HP2001-0044), Xunta de Galicia (PGIDT01PXI10506PN, PGIDIT02PXIB30501PR) and Universidade da Coruña.

with derivational morphology.

In this process, the first step consists of tagging the document. Document processing starts by applying our linguistically-motivated preprocessor module [8, 2], performing tasks such as format conversion, tokenization, sentence segmentation, morphological pretagging, contraction splitting, separation of enclitic pronouns from verbal stems, expression identification, numeral identification and proper noun recognition. It is interesting to remark that classical techniques do not deal with many of these phenomena, resulting in wrong simplifications during conflation process.

The output of the preprocessor is taken as input by the tagger-lemmatizer. Although any kind of tagger could be applied, in our system we have used a second order Markov model for part-of-speech tagging. The elements of the model and the procedures to estimate its parameters are based on Brant's work [3], incorporating information from external dictionaries [9] which are implemented by means of numbered minimal acyclic finite-state automata [7].

Once text has been tagged, the lemmas of the content words (nouns, verbs, adjectives) are extracted to be indexed. In this way we are solving the problems derived from inflection in Spanish and, as a result, recall is increased. With regard to computational cost, the running cost of a lemmatizer-disambiguator is linear in relation to the length of the word, and cubic in relation to the size of the tagset, which is a constant. As we only need to know the grammatical category of the word, the tagset is small and therefore the increase in cost with respect to classical approaches (stemmers) becomes negligible.

Now inflectional variation has been solved, the next logical step is to solve the problems caused by derivational morphology. 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 of word formation. We have considered the derivational morphemes, the allomorphic variants of such morphemes and the phonological conditions they must satisfy, to automatically generate the set of morphological families from a large lexicon of Spanish words [18]. The resulting morphological families can be used as a kind of advanced and linguistically motivated stemmer for Spanish, where every lemma is substituted by a fixed representative of its morphological family. Since the set of morphological families is generated statically, there is no increment in the running cost.

3 Using synonymy to expand queries

The use of synonymy relations in the task of automatic query expansion is not a new subject, but the approaches presented until now do not assign a weight to the degree of synonymy that exists between the original terms present in the query and those produced by the process of expansion [10]. Nevertheless, our system does have access to this information, so a threshold of synonymy can be set in order to control the degree of query expansion.

The most frequent definition of synonymy conceives it as a relation between two expressions with identical or similar meaning. The controversy of understanding synonymy as a precise question or as an approximate question, i.e. as a question of identity or as a question of similarity, has existed from the beginning of the study of this semantic relation. In our system, synonymy is understood as a gradual relation between words. In order to calculate the degree of synonymy, we use the *Jaccard's coefficient* as measure of similarity applied on the sets of synonyms provided by a dictionary of synonyms for each of its entries [5]. Given two sets X and Y, their *similarity* is measured as:

$$sm(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Let us consider a word w with m_i possible meanings, and another word w' with m_j possible meanings, where $dc(w,m_i)$ represents the function that gives us the set of synonyms provided by the dictionary for every entry w in the concrete meaning m_i. The degree of synonymy of w and w' in the meaning m_i of w is calculated as $dg(w,m_i,w') = \max_j \; sm[dc(w,m_i),dc(w',m_j)]$. Furthermore, by calculating $k = \arg\max_j \; sm[dc(w,m_i),dc(w',m_j)]$ we obtain in m_k the meaning of w' closest to the meaning m_i of w.

4 Extracting dependencies between words by means of a shallow parser

Our system is not only able to process the content of the document at word level, it can also process its syntactic structure. For this purpose, a parser module obtains from the tagged document the *head-modifier* pairs corresponding to the most relevant syntactic dependencies: *noun-modifier*, relating the head of a noun phrase with the head of a modifier; *subject-verb*, relating the head of the subject with the main verb of the clause; and *verb-complement*, relating the main verb of the clause with the head of a complement.

The kernel of the grammar used by this shallow parser is inferred from the basic trees corresponding to noun phrases[1] and their syntactic and morpho-syntactic variants [11, 17]:

- *Syntactic variants* result from the inflection of individual words and from modifying the syntactic structure of the original noun phrase by means of:
  - *Synapsy:* it corresponds to a change of preposition or the addition or removal of a determiner.
    una caída de ventas (a drop in sales)
  - *Substitution:* it consists of employing modifiers to make a term more specific.
    una caída inusual de ventas (an unusual drop in sales)
  - *Permutation:* this refers to the permutation of words around a pivot element.
    una inusual caída de ventas (an unusual drop in sales)
  - *Coordination:* this consists of employing coordinating constructions (copulative or disjunctive) with the modifier or with the modified term.
    una inusual caída de ventas y de beneficios (an unusual drop in sales and profits)
- *Morpho-syntactic variants* differ from syntactic variants in that at least one of the content words of the original noun phrase is transformed into another word derived from the same morphological stem.
  las ventas han caído (sales have dropped)

We must remark that syntactic variants involve inflectional morphology but not derivational morphology, whereas morpho-syntactic variants involve both inflectional and derivational morphology. In addition, syntactic variants have a very restricted scope (the noun phrase) whereas morpho-syntactic variants can span a whole sentence, including a verb and its complements.

Once the basic trees of noun phrases and their variants have been established, they are compiled into a set of regular expressions, which are matched against the tagged document in order to extract its dependencies in the form of pairs which are used as index terms after conflating their components through morphological families, as is described in [17]. In this way, we are identifying dependency pairs through simple pattern matching over the output of the tagger-lemmatizer, solving the problem by means of finite-state techniques, leading to a considerable reduction of the running cost.

5 Non-official experiments with CLEF 2001 queries

The Spanish corpus was incorporated in CLEF 2001 [16], but the techniques proposed in this paper have been integrated very recently and so we could not participate in that edition. Nevertheless, we consider interesting to present the results of some non-official experiments performed with the set of queries of CLEF 2001[2].

The Spanish CLEF corpus is formed by 215,738 documents corresponding to the news provided by EFE, a Spanish news agency, in 1994. Documents are formatted in SGML, with a total size of 509 Megabytes. After deleting SGML tags, the size of the text corpus is reduced to 438 Megabytes. Each query consists of three fields: a brief title statement, a one-sentence description, and a more complex narrative specifying the relevance assessment criteria. In these experiments, we have employed the three fields to build the final query submitted to the system. For linguistically-motivated indexing techniques, the terms contained in the title section are given the double of importance with respect to description and narrative.

The techniques proposed in this article are independent of the indexing engine we choose to use. This is because we first conflate the 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 [3] [19]. The results we show here have been obtained with SMART, using the `ltc-lnc` weighting scheme [4], without relevance feedback.

We have compared the results obtained by four different indexing methods:

[1] At this point we will take as example the noun phrase *una caída de las ventas* (a drop in the sales).

[2] We have also tested some of the techniques proposed in this article over our own, non standard, corpus, formed by 21,899 news articles (national, international, economy, culture,...). Results are reported in [19].

[3] Indexing model, ranking algorithm, etc.

Table 1: Number of index terms extracted from the CLEF corpus

	plain text	***stm***	***lem***	***fam***	***f-sdp***
Total	68,530,085	33,712,903	33,158,582	33,158,582	58,497,396
Unique	529,914	345,435	388,039	384,003	5,129,665

Table 2: CLEF 2001: performance measures

	stm	***lem***	***fam***	***f-sdp***
Documents retrieved	49,000	49,000	49,000	49,000
Relevant documents retrieved (2694 expected)	2,576	2,554	2,563	2,565
R-precision	0.4787	0.4809	0.4814	0.4692
Average precision per query	0.4915	0.4749	0.4843	0.4669
Average precision per relevant docs	0.5561	0.5521	0.5492	0.5189
11-points average precision	0.4976	0.4864	0.4927	0.4799

- Stemming text after eliminating stopwords (*stm*). In order to apply this technique, we have tested several stemmers for Spanish. Finally, the best results we obtained were for the stemmer used by the open source search engine Muscat[4], based on Porter's algorithm [1].
- Conflation of content words via lemmatization (*lem*), i.e. each form of a content word is replaced by its lemma. This kind of conflation takes only into account inflectional morphology.
- Conflation of content words by means of morphological families (*fam*), i.e. each form of a content word is replaced by the representative of its morphological family. This kind of conflation takes into account both inflectional and derivational morphology.
- Text conflated by means of the combined use of morphological families and syntactic dependency pairs (*f-sdp*).

The methods *lem*, *fam*, and *f-sdp* are linguistically motivated. Therefore, they are able to deal with some complex linguistic phenomena such as clitic pronouns, contractions, idioms, and proper name recognition. In contrast, the method *stm* works simply by removing a given set of suffixes, without taking into account such linguistic phenomena, yielding incorrect conflations that introduce noise in the system. For example, clitic pronouns are simply considered a set of suffixes to be removed. Moreover, the employment of finite-state techniques in the implementation of our methods let us to reduce their computational cost, making possible their application in practical environments.

Table 1 shows the statistics of the terms that compose the 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. Table 2 shows performance measures as defined in the standard `trec_eval` program. The monolingual Spanish task in 2001 considered a set of 50 queries, but for one query any relevant document exists in the corpus, and so the performance measures are computed over 49 queries. Table 3 shows in its left part the precision attained at the 11 standard recall levels. We can observe that linguistically motivated indexing techniques beats *stm* for low levels of recall. This fact means that more highly relevant documents are placed in the top part of the ranking list applying these techniques. As a complement, the right part of Table 3 shows the precision computed at N seen documents.

The results of our experiments seems to be consistent with the results obtained for English and Germanic languages by other IR systems based on NLP techniques [12, 13, 14, 15]. As in [14], syntax does not improve average precision, but is the best technique for low levels of recall. A similar conclusion can be extracted from the work of [12] on Dutch texts, where syntactic methods only beats statistical ones at low levels of recall. Our results with respect to syntactic dependency pairs seem to be better that those of Perez-Carballo and Strzalkowski [15]. It

[4]Currently, Muscat is not an open source project, and the web site `http://open.muscat.com` used to download the stemmer is not operating. Information about a similar stemmer for Spanish (and other European languages) can be found at `http://snowball.sourceforge.net/spanish/stemmer.html`.

Table 3: CLEF 2001: average precision at 11 standard recall levels and at N seen documents

Recall	Precision				N	Precision			
	stm	***lem***	***fam***	***f-sdp***		***stm***	***lem***	***fam***	***f-sdp***
0.00	0.8426	0.8493	0.8518	0.8658	5	0.6122	0.6204	0.6367	0.5918
0.10	0.7539	0.7630	0.7491	0.7422	10	0.5551	0.5245	0.5429	0.5143
0.20	0.6971	0.6738	0.6895	0.6766	15	0.5075	0.4871	0.4925	0.4612
0.30	0.6461	0.6117	0.6312	0.6047	20	0.4735	0.4500	0.4510	0.4398
0.40	0.5669	0.5589	0.5656	0.5305	30	0.4238	0.4136	0.4095	0.3980
0.50	0.5013	0.4927	0.4979	0.4687	100	0.2827	0.2759	0.2769	0.2661
0.60	0.4426	0.4209	0.4252	0.4211	200	0.1893	0.1903	0.1877	0.1813
0.70	0.3832	0.3636	0.3641	0.3444	500	0.0979	0.0969	0.0970	0.0952
0.80	0.3221	0.3080	0.3109	0.2941	1000	0.0526	0.0521	0.0523	0.0523
0.90	0.2140	0.2109	0.2221	0.2113					
1.00	0.1037	0.0974	0.1126	0.1194					

is difficult to know if this improvement is due to a more accurate extraction of pairs or due to differences between Spanish and English constructions.

6 Experiments with CLEF 2002 queries

6.1 The uppercase-to-lowercase module

An important characteristic of IR test collections that may have a considerable impact on the performance of linguistically motivated indexing techniques is the large number of typographical errors present in documents, as have been reported, in the case of the Spanish CLEF corpus, by [6]. In particular, titles of news and subsections are generally written in capital letters without accents. We must take into account that these titles are usually very indicative of the topic of the document.

For CLEF 2002 experiments, we have incorporated an *uppercase-to-lowercase* module to our system to process uppercase sentences, converting them to lowercase and restoring the existent diacritics when necessary. Other approaches, such as [20], deal with documents where absolutely all diacritics have been eliminated. Nevertheless, our situation is different, because the main of the document is written lowercase and preserves their diacritics, only some sentences are written in capital letters; moreover, for our purposes we only need the grammatical category and lemma of the word, not the form.

So, we can employ the lexical context of an uppercase sentence, either forms and lemmas, to recover this lost information. The first step of this process is to identify the uppercase phrases. We consider that a sequence of words form an *uppercase phrase*, when it consists of three or more words written in capital letters and at least three of them have more than three characters. For each of these uppercase phrases we do the following:

1. We obtain its surrounding context.
2. For each of the words in the phrase:
   (a) We examine the context looking for entries with the same flattened form [5]. Each of these words become candidates.
   (b) If candidates are found, the most numerous is chosen, and in case of existing a draw, the closest to the phrase is chosen.
   (c) If no candidates are found, the lexicon is examined:
      i. We obtain from the lexicon all entries with the same flattened form, grouping them according to their category and lemma (we are not interested in the form, just in the category and the lemma of the word).
      ii. If no entries are found, we keep the actual tag and lemma.
      iii. If only one entry is found, we choose that one.
      iv. If more than one entry is found, we choose the most numerous in the context (according to the category and the lemma). Again, in case of existing a draw, we choose the closest to the sentence.

Sometimes, some words of the uppercase phrase preserve some of their diacritics, for example the ˜ of the Ñ. In this situations, the candidates from the context or the lexicon must observe this restriction.

[5] That is, after both words been converted to lowercase, and after eliminating all diacritics from them

Table 4: CLEF 2002: performance measures

	TD***lem***	TDN***lem***	TDN***syn***	TDN***pds***
Documents retrieved	50,000	50,000	50,000	50,000
Relevant documents retrieved (2854 expected)	2,495	2,634	2,632	2,624
R-precision	0.3697	0.4466	0.4438	0.3983
Average precision per query	0.3608	0.4448	0.4423	0.4043
Average precision per relevant docs	0.3971	0.4665	0.4613	0.4472
11-points average precision	0.3820	0.4630	0.4608	0.4205

Table 5: CLEF 2002: average precision at 11 standard recall levels and at N seen documents

Recall	Precision				N	Precision			
	TD***lem***	TDN***lem***	TDN***syn***	TDN***pds***		TD***lem***	TDN***lem***	TDN***syn***	TDN***pds***
0.00	0.7173	0.8065	0.8052	0.7732	5	0.4600	0.5480	0.5440	0.5680
0.10	0.6226	0.7311	0.7199	0.6785	10	0.4540	0.5280	0.5280	0.5280
0.20	0.5587	0.6699	0.6637	0.6158	15	0.4133	0.5000	0.4933	0.4880
0.30	0.4786	0.5788	0.5812	0.5278	20	0.4050	0.4760	0.4690	0.4510
0.40	0.4375	0.5360	0.5355	0.4816	30	0.3693	0.4327	0.4280	0.3960
0.50	0.3877	0.4645	0.4619	0.4263	100	0.2348	0.2636	0.2642	0.2474
0.60	0.3154	0.3971	0.4004	0.3598	200	0.1645	0.1831	0.1825	0.1780
0.70	0.2673	0.3560	0.3558	0.3009	500	0.0892	0.0952	0.0952	0.0933
0.80	0.2061	0.2769	0.2704	0.2307	1000	0.0499	0.0527	0.0526	0.0525
0.90	0.1480	0.1883	0.1908	0.1624					
1.00	0.0627	0.0877	0.0837	0.0689					

6.2 Results

We have compared the results obtained by four different indexing methods:

- TD*lem*: Conflation of content words via lemmatization, i.e. each form of a content word is replaced by its lemma. This kind of conflation takes only into account inflectional morphology. The query is formed by the set of meaning lemmas present in title and description.

- TDN*lem*: The same as before, but the query also includes the set of meaning lemmas obtained from the narrative. Both this method and the previous one correspond to the *lem* indexing method referred in Section 5.

- TDN*syn*: Conflation of content words via lemmatization and expansion of queries by means of synonymy. We have considered that two words are synonyms if their similarity measure is greater or equal to 0.80. The query is formed by the set of meaning lemmas present in title, description and narrative, but only the title and description field of each query have been expanded using synonyms.

- TDN*pds*: Text conflated by means of the combined use of morphological families and syntactic dependency pairs. The query is formed by the union of the set of representatives of the morphological families corresponding to the content words and the set of dependency pairs extracted from the title, description and narrative fields. It corresponds to the *f-sdp* indexing method referred in Section 5.

Except for the first method, the terms extracted from the title section are given the double of importance with respect to description and narrative.

According to Tables 4 and 5, the lemmatization method (TDN*lem*) seems to be the best option. The expansion through synonymy (TDN*syn*) does not improve the results obtained, perhaps because the expansion is *total*, that is, *all* synonyms of *all* terms of the query are employed, introducing too much noise. In the case of the employment of syntactic dependency pairs (TDN*pds*), the results are worse than for CLEF 2001 queries. This may be simply due to the different set of queries employed, but after comparing the results of each particular query with lemmatization, it may be concluded that the more accurate is the complex term with respect to its constituting simple terms, the more the results improve, as in the case of *estadísticas de divorcio* (divorce statistics) in the 115th query.

These results, together with the previous ones obtained for CLEF 2001 queries, suggest that mere lemmatization is a good starting point. It may be investigated whether this initial search should be followed by a relevance feedback process based on the expansion of the synonyms of the most relevant terms of the most relevant documents to minimize the noise. Another alternative to study for postprocessing consists on the reranking of the results by means of syntactic information obtained in form of syntactic dependency pairs.

References

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*. Addison-Wesley, Harlow, England, 1999.

[2] Fco. Mario Barcala, Jesús Vilares, Miguel A. Alonso, Jorge Graña, and Manuel Vilares. Tokenization and proper noun recognition for information retrieval. In *3rd International Workshop on Natural Language and Information Systems (NLIS 2002), September 2-3, 2002. Aix-en-Provence, France*, Los Alamitos, California, USA, September 2002. IEEE Computer Society Press.

[3] Thorsten Brants. TNT - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP'2000)*, Seattle, 2000.

[4] Chris Buckley, James Allan, and Gerard Salton. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In D. K. Harman, editor, *NIST Special Publication 500-215: The Second Text REtrieval Conference (TREC-2)*, pages 45–56, Gaithersburg, MD, USA, 1993.

[5] Santiago Fernández, Jorge Graña, and Alejandro Sobrino. A Spanish e-dictionary of synonyms as a fuzzy tool for information retrieval. In *Actas de las I Jornadas de Tratamiento y Recuperación de Información (JOTRI 2002)*, León, Spain, September 2002.

[6] Carlos G. Figuerola, Raquel Gómez, Angel F. Zazo Rodríguez, and José Luis Alonso Berrocal. Stemming in Spanish: A first approach to its impact on information retrieval. In Carol Peters, editor, *Working notes for the CLEF 2001 workshop*, Darmstadt, Germany, September 2001.

[7] Jorge Graña, Fco. Mario Barcala, and Miguel A. Alonso. Compilation methods of minimal acyclic automata for large dictionaries. In Bruce W. Watson and Derick Wood, editors, *Proc. of the 6th Conference on Implementations and Applications of Automata (CIAA 2001)*, pages 116–129, Pretoria, South Africa, July 2001.

[8] Jorge Graña, Fco. Mario Barcala, and Jesús Vilares. Formal methods of tokenization for part-of-speech tagging. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 240–249. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[9] Jorge Graña, Jean-Cédric Chappelier, and Manuel Vilares. Integrating external dictionaries into stochastic part-of-speech taggers. In *Proceedings of the Euroconference Recent Advances in Natural Language Processing (RANLP 2001)*, pages 122–128, Tzigov Chark, Bulgaria, 2001.

[10] Jane Greenberg. Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science and Technology*, 52(5):402–415, 2001.

[11] Christian Jacquemin and Evelyne Tzoukermann. NLP for term variant extraction: synergy between morphology, lexicon and syntax. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, volume 7 of *Text, Speech and Language Technology*, pages 25–74. Kluwer Academic Publishers, Dordrecht/Boston/London, 1999.

[12] Wessel Kraaij and Renée Pohlmann. Comparing the effect of syntactic vs. statistical phrase indexing strategies for Dutch. In Christos Nicolaou and Constantine Stephanidis, editors, *Research and Adavanced Technology for Digital Libraries*, volume 1513 of *Lecture Notes in Computer Science*, pages 605–614. Springer-Verlag, Berlin/Heidelberg/New York, 1998.

[13] Byung-Kwan Kwak, Jee-Hyub Kim, Geunbae Lee, and Jung Yun Seo. Corpus-based learning of compound noun indexing. In J. Klavans and J. Gonzalo, editors, *Proc. of the ACL'2000 workshop on Recent Advances in Natural Language Processing and Information Retrieval*, Hong Kong, October 2000.

[14] Markus Mittendorfer and Werner Winiwarter. Exploiting syntactic analysis of queries for information retrieval. *Data & Knowledge Engineering*, 2002.

[15] Jose Perez-Carballo and Tomek Strzalkowski. Natural language information retrieval: progress report. *Information Processing and Management*, 36(1):155–178, 2000.

[16] Carol Peters, editor. *Results of the CLEF 2001 Cross-Language System Evaluation Campaign. Working Notes for the CLEF 2001 Workshop*, Darmstadt, Germany, September 2001.

[17] Jesús Vilares, Fco. Mario Barcala, and Miguel A. Alonso. Using syntactic dependency-pairs conflation to improve retrieval performance in Spanish. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science,*, pages 381–390. Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[18] Jesús Vilares, David Cabrero, and Miguel A. Alonso. Applying productive derivational morphology to term indexing of Spanish texts. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2004 of *Lecture Notes in Computer Science*, pages 336–348. Springer-Verlag, Berlin-Heidelberg-New York, 2001.

[19] Jesús Vilares, Manuel Vilares, and Miguel A. Alonso. Towards the development of heuristics for automatic query expansion. In Heinrich C. Mayr, Jiri Lazansky, Gerald Quirchmayr, and Pavel Vogel, editors, *Database and Expert Systems Applications*, volume 2113 of *Lecture Notes in Computer Science*, pages 887–896. Springer-Verlag, Berlin-Heidelberg-New York, 2001.

[20] David Yarowsky. A comparison of corpus-based techniques for restoring accents in Spanish and French text. In *Natural Language Processing Using Very Large Corpora*, pages 99–120. Kluwer Academic Publishers, 1999.