On the Usefulness of Extracting Syntactic Dependencies for Text Indexing*

Miguel A. Alonso[1], Jesús Vilares[1,2], and Víctor M. Darriba[2]

[1] Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 La Coruña, Spain
alonso@udc.es jvilares@mail2.udc.es
http://coleweb.dc.fi.udc.es/
[2] Escuela Superior de Ingeniería Informática, Universidade de Vigo
Campus de As Lagoas, 32004 Orense, Spain
jvilares@uvigo.es darriba@ei.uvigo.es

Abstract. In recent years, there has been a considerable amount of interest in using Natural Language Processing in Information Retrieval research, with specific implementations varying from the word-level morphological analysis to syntactic parsing to conceptual-level semantic analysis. In particular, different degrees of phrase-level syntactic information have been incorporated in information retrieval systems working on English or Germanic languages such as Dutch. In this paper we study the impact of using such information, in the form of syntactic dependency pairs, in the performance of a text retrieval system for a Romance language, Spanish.

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. More elaborated index terms can be created by combining two or more content words (nouns, verbs and adjectives) in a *multi-word term* [11, 6, 12]. Most techniques for extracting multi-word terms rely on statistics [7] or simple pattern matching [12], instead of considering the structural relations among the words that form a sentence. In this paper, we propose to use practical, finite-state, Natural Language Processing (NLP) techniques to extract such multi-word terms in the form of pairs of words related by some kind of syntactic dependency.

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

2 Extraction of Syntactic Dependencies

Given a stream of tagged words, we want to obtain the *head-modifier* pairs corresponding to the most relevant syntactic dependencies [5]: *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. It has to be noted that while the head-modifier relation may suggest semantic dependence, what we obtain here is strictly syntactic, even though the semantic relation is what we are really after [16].

The kernel of the grammar used by our shallow parser has been inferred from the basic trees corresponding to noun phrases and their syntactic and morpho-syntactic variants [11]. *Syntactic variants* result from the inflection of individual words and from modifying the syntactic structure of the original noun phrase. *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. At this point we must recall that inflectional morphemes represent grammatical concepts such as gender, person, mood, or time and tense. On the other hand, derivational morphemes effect a semantic change on the base, often also effecting a change of syntactic class. We define a *morphological family* as the set of words obtained from the same morphological root through derivation mechanisms, such as prefixation, emotive suffixation, non-emotive suffixation, back formation and parasynthesis. A system for the automatic generation of morphological families has been described in [18].

2.1 Syntactic Variants

The example of Fig. 1 shows the basic structure of a noun phrase and some of its possible syntactic variants, together with the syntactic dependencies they contain. Such variants were obtained by applying to the source phrase *una caída de las ventas* (a drop in the sales) the following mechanisms [11]:

- *Synapsy:* a unary construction which corresponds to a change of preposition or the addition or removal of a determiner.
- *Substitution:* it consists of employing modifiers to make a term more specific.
- *Permutation:* this refers to the permutation of words around a pivot element.
- *Coordination:* this consists of employing coordinating constructions (copulative or disjunctive) with the modifier or with the modified term.

Symbols A, C, D, N, P, V and W are the part-of-speech labels that denote adjectives, coordinating conjunctions, determiners, nouns, prepositions, verbs and adverbs, respectively. In addition, we have conflated each word in a dependency pair by replacing it with an identifier of its morphological family (actually, one of the words in such family, its *representative*).

The structures and syntactic dependencies corresponding to all the syntactic variants shown in Fig. 1 are embedded in the syntactic pattern shown in Fig. 2

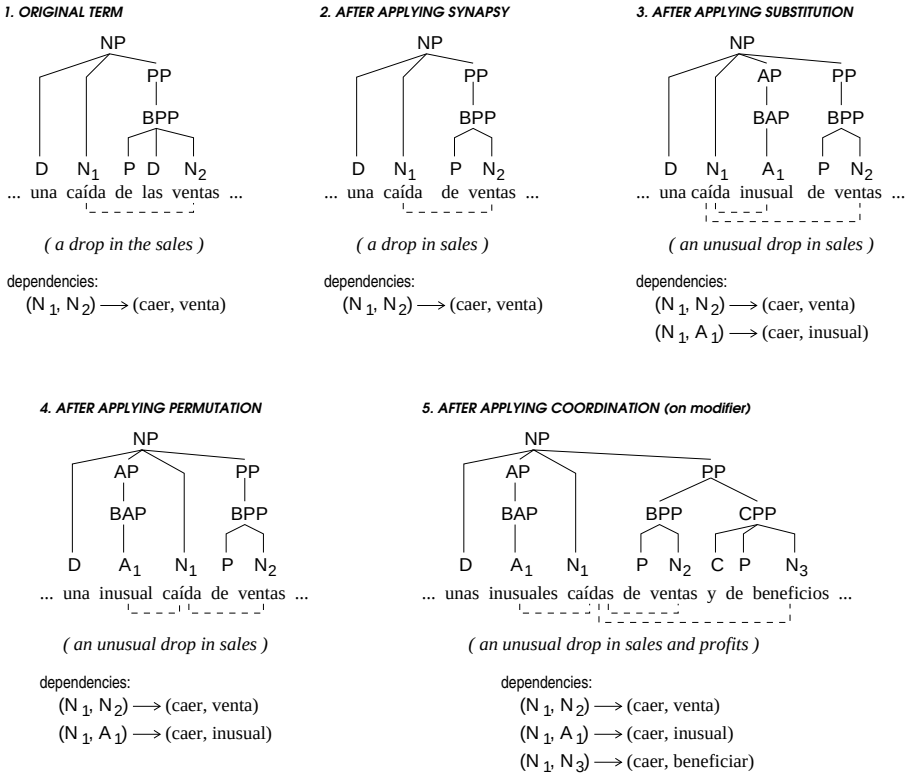

Fig. 1. Syntactic variants of *una caída de las ventas* (a drop in the sales)

and its twin, in which the *adjective phrase* postposed to the head name (shaded) is placed before such name. We have followed the finite-state shallow parsing approach used in successful information extraction systems [1, 9, 10], instead of the full parsing approach applied by some information retrieval systems [16]. In our particular example, the syntactic pattern is translated into the following regular expression, $D?\ N_1\ (W?\ A_1\ (C\ W?\ A_i))?\ P\ D?\ N_2\ A_2?\ (C\ P\ D?\ N_3\ A_3?)?$, keeping its associated syntactic dependency pairs. In this way, we can identify and extract multi-word index terms through simple pattern matching over the output of the tagger/lemmatizer, dealing with the problem from a surface processing approach at lexical level, leading to a considerable reduction of the running cost.

2.2 Morpho-Syntactic Variants

Morpho-syntactic variants are classified according to the nature of the derivational transformations applied to their words:

– *Iso-categorial:* morphological derivation process does not change the category of words, but only transforms one noun syntagma into another. There are two possibilities: noun-to-noun and adjective-to-adjective.

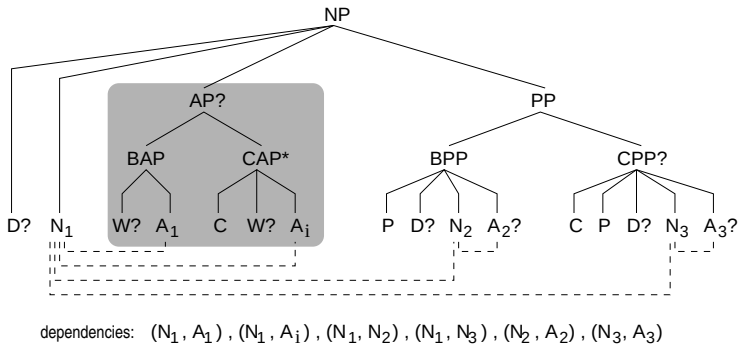

Fig. 2. Syntactic pattern for the syntactic variants of *una caída de las ventas*

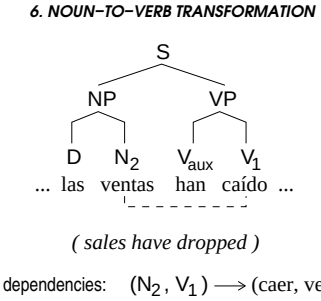

Fig. 3. Morpho-syntactic variant of *una caída de las ventas* (a drop in the sales)

– *Hetero-categorial:* morphological derivation does result in a change of the category of a word. There are also two possibilities: noun-to-verb and noun-to-adjective.

Figure 3 shows the structure of the sentence *las ventas han caído* (sales have dropped), a morpho-syntactic variant of *una caída de las ventas* (a drop in the sales) that involves a noun-to-verb transformation: *caída* (drop) becomes *caer* (to drop). We can observe that the words in the dependency pair extracted are the nouns *caída* (drop) and *ventas* (sales) in the source phrase, and the verb *caer* (to drop) and the noun *ventas* (sales) in its variant; therefore, different syntactic dependency pairs would be obtained. However, there exist a derivational relation between *caer* and *caída* and so, by employing morphological families to conflate the components of syntactic dependency pairs, we are able to obtain the same pair for both the source phrase and the variant. Therefore, common information to both the variant and the original term is conflated in the same way, which is the objective pursued [2].

Finally, 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 vari-

ants can span a whole sentence, including a verb and its complements, as in the case of Fig. 3.

3 Evaluation

The lack of a standard evaluation corpus has been a great handicap for the development of IR research in Spanish.[1] This situation is changing due to the incorporation in CLEF-2001 [17] of a Spanish corpus which is expected to become a standard. The techniques proposed in this paper have been integrated very recently, and therefore we could not participate in the 2001 edition, but we have joined CLEF competition in 2002, a fact that has allowed us to access the document collection and queries of the previous edition. Thus, the results reported in this section correspond to non-official experiments.[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. We have employed the three fields to build the final query submitted to the system.

The conflation of multi-word terms by means of the extraction of syntactic dependency pairs from queries and documents is independent of the indexing engine and so, any standard text indexing engine may be employed. Nevertheless, each engine will behave according to its own characteristics, such as indexing model, ranking algorithm, etc. [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:

- Stemmed 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[3], based on Porter's algorithm [3].
- Conflation of content words (nouns, adjectives and verbs) 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.

[1] The test collection used in the Spanish track of TREC-4 (1995) and TREC-5 (1996), formed by news articles written in Mexican-Spanish, is no longer freely available.

[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,...) with an average length of 447 words, considering a set of 14 natural language queries with an average length of 7.85 words per query, 4.36 of which were content words. Results are reported in [19].

[3] 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 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. Performance measures

	stm	***lem***	***fam***	***f-sdp***
Documents retrieved	49,000	49,000	49,000	49,000
Relevant documents retrieved	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

– 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.

Table 3. 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					

4 Conclusion

In this article we have studied how the extraction of head-modifier pairs impacts the performance of text retrieval systems. Albeit our scheme is oriented towards the indexing of Spanish texts, it is also a proposal of a general architecture that can be applied to other languages with slight modifications. We have tested our approach with the CLEF 2001 collection of documents and queries, and the results of our experiments are consistent with the results obtained for English and Germanic languages by other IR systems based on NLP techniques [15, 13, 14, 16]. As in [15], 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 [13] 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 [16]. It 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.

An important characteristic of the CLEF collection that can 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 [8]. In particular, titles of the news (documents) are in capital letters without accents. We must take into account that the title of a news article is usually very indicative of its topic.

References

1. C. Aone, L. Halverson, T. Hampton, and M. Ramos-Santacruz. SRA: Description of the IE^2 system used for MUC-7. In *Proc. of the MUC-7*, 1998.
2. A. Arampatzis, T. van der Weide, C. Koster, and P. van Bommel. Linguistically motivated information retrieval. In *Encyclopedia of Library and Information Science*. Marcel Dekker, Inc., New York and Basel, 2000.
3. R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. Addison-Wesley, Harlow, England, 1999.

4. C. Buckley, J. Allan, and G. Salton. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In D. K. Harman, editor, *Proc. of TREC-2*, pages 45–56, Gaithersburg, MD, USA, 1993.
5. J. Carrol, T. Briscoe, and A. Sanfilippo. Parser evaluation: a survey and a new proposal. In *Proc. of LREC'98*, pages 447–454, Granada, Spain, 1998.
6. 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.
7. J. L. Fagan. Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *Proc. of SIGIR'87*, pages 91–101, 1987.
8. C. G. Figuerola, R. Gómez, A. F. Zazo, and J. L. Alonso. Stemming in Spanish: A first approach to its impact on information retrieval. In [17].
9. R. Grishman. The NYU system for MUC-6 or where's the syntax? In *Proc. of MUC-6*. Morgan Kaufmann Publishers, 1995.
10. J. R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*. MIT Press, Cambridge, MA, USA, 1997.
11. C. Jacquemin and E. Tzoukermann. NLP for term variant extraction: synergy between morphology, lexicon and syntax. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25–74. Kluwer Academic Publishers, Dordrecht/Boston/London, 1999.
12. 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.
13. W. Kraaij and R. Pohlmann. Comparing the effect of syntactic vs. statistical phrase indexing strategies for Dutch. In C. Nicolaou and C. Stephanidis, editors, *Research and Adavanced Technology for Digital Libraries*, volume 1513 of *LNCS*, pages 605–614. Springer-Verlag, Berlin/Heidelberg/New York, 1998.
14. B.-K. Kwak, J.-H. Kim, G. Lee, and J. Y. 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.
15. M. Mittendorfer and W. Winiwarter. Exploiting syntactic analysis of queries for information retrieval. *Data & Knowledge Engineering*, 2002.
16. J. Perez-Carballo and T. Strzalkowski. Natural language information retrieval: progress report. *Information Processing and Management*, 36(1):155–178, 2000.
17. C. Peters, editor. *Working Notes for the CLEF 2001 Workshop*. Darmstadt, Germany, 2001. Available at `http://www.clef-campaign.org`.
18. J. Vilares, D. Cabrero, and M. 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 *LNCS*, pages 336–348. Springer-Verlag, Berlin-Heidelberg-New York, 2001.
19. J. Vilares, M. Vilares, and M. 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*, volume 2113 of *LNCS*, pages 887–896. Springer-Verlag, Berlin-Heidelberg-New York, 2001.