

# Abstract

This PhD. thesis belongs to both the fields of Natural Language Processing, the area of science and technology which deals with the automatic processing of natural or human language, and Information Retrieval, whose goal consists of identifying the documents in a collection which are relevant to a given information need of the user.

Conventional Information Retrieval systems employ statistical techniques based on the distribution of terms through the document and the collection to calculate the relevance of a document. Nevertheless, since an Information Retrieval process requires the system to understand, in some way, the content of the document, such a task can be also viewed as a Natural Language Processing task. This reasoning is supported by the fact that the major problem of Information Retrieval is linguistic variation of language, namely that the same concept can be expressed in different ways by changing the expression.

The aim of this PhD. thesis is the development of base technology for Natural Language Processing and the study of the viability of its application in Spanish Information Retrieval systems. Although similar works have been done for other languages, with English at the fore, Spanish has stayed in the background. Moreover, the greater linguistic complexity of Spanish with respect to English does not allow a direct extrapolation of the results obtained for English, demanding, in this way, that Spanish has its own experiments.

Furthermore, we have had to face one of the main problems in Natural Language Processing research for Spanish, the lack of freely available linguistic resources. The solution for minimizing this problem consists in restricting the complexity of the solutions proposed, by focusing on the employment of lexical information, which is easier to obtain. The fact of limiting the complexity of the approaches proposed allows the techniques developed to be easily adaptable to other languages with similar characteristics and behavior, resulting in a general architecture which can be applied to other languages by introducing the appropriate modifications, as in the case of Galician, Portuguese or Catalan, for example. On the other hand, in order to minimize the computational cost of our approaches for their application in practical environments, finite-state technology has been widely used. The general architecture proposed is described below.

Firstly, an advanced linguistically-based preprocessor has been developed for tokenization and segmentation of Spanish texts, mainly orientated to robust part-of-speech tagging of Spanish, but also applicable to other analysis tasks. Concrete modifications made for its application in Information Retrieval are also discussed.

At inflectional level, the employment of lemmatization-disambiguation techniques for single word term conflation has been studied, using as index terms the lemmas of content words — nouns, adjectives and verbs.

At derivational level, a tool for the automatic generation of morphological families —sets of derivationally related words that share the same root— has been developed for its employment in single word term conflation. Nevertheless, this approach is not wholly exempt from the problems caused by the noise introduced because of overgeneration during the generation of families.

At syntactic level, an approach based on the use of syntactic dependencies as complex index

terms has been tested in order to obtain more precise index terms and to manage syntactic and morpho-syntactic variation. These dependencies are obtained through shallow parsing by employing two parsers developed for this purpose: PATTERNS, based on patterns, and CASCADE, based on a cascade of finite-state transducers. The use of different sources of syntactic information, queries or documents, has been also studied—the latter being more effective—, as has the restriction of the dependencies employed to only those obtained from noun phrases in order to reduce costs.

Finally, also at syntactic level, a new pseudo-syntactic approach, which employs a retrieval model based on a similarity measure computed as a function of the distance between terms is used for reranking the documents obtained by a classical approach based on the indexing of lemmas. Two different approaches are proposed, the first one based on the mere reranking of the documents according to the new retrieval model, and the second one based on a new data fusion method which employs set intersection, this second approach being more fruitful.