Abstract. In this our second participation in the CLEF Spanish monolingual track, we have continued applying Natural Language Processing techniques for single word and multi-word term conflation. Two different conflation approaches have been tested. The first approach is based on the lemmatization of the text in order to avoid inflectional variation. Our second approach consists of the employment of syntactic dependencies as complex index terms, in an attempt to solve the problems derived from syntactic variation and, in this way, to obtain more precise terms. Such dependencies are obtained through a shallow parser based on cascades of finite-state transducers.

1 Introduction

In Information Retrieval (IR) systems, the correct representation of a document through an accurate set of index terms is the basis for obtaining a good performance. If we are not able to both extract and weight appropriately the terms which capture the semantics of the text, this shortcoming will have an effect on all the subsequent processing.

In this context, one of the major limitations we have to deal with is the linguistic variation of natural languages [5], particularly when processing documents written in languages with more complex morphologic and syntactic structures than those present in English, as in the case of Spanish. When managing this type of phenomena, the employment of Natural Language Processing (NLP) techniques becomes feasible. This has been our working hypothesis since our research group, the COLE Group, started its work on Spanish Information Retrieval.

As in our first participation in CLEF [23], our main premise is the search for simplicity, motivated by the lack of available linguistic resources for Spanish such as large tagged corpora, treebanks or advanced lexicons. This work is thus a continuation and refinement of the previous work presented in CLEF 2002, but
centered this time on the employment of lemmatization for solving the *inflectional variation* and the employment of syntactic dependencies for solving the *syntactic variation*.

This article is outlined as follows. Section 2 describes the techniques used for single word term conflation. Section 3 introduces our approach for dealing with syntactic variation through shallow parsing. Official runs are presented and discussed in Section 4. Next, Section 5 describes the set of experiments performed after our participation in the workshop, in an attempt to eliminate some of the drawbacks detected. Finally, our conclusions and future developments are presented in Section 6.

### 2 Single Word Term Conflation

Our proposal for single word term conflation continues to be based on exploiting the lexical level in two phases: first, by solving the *inflectional variation* through lemmatization, and second, by solving the *derivational morphology* through the employment of morphological families.

The process followed for single word term conflation starts by tagging the document. The first step consists in applying our linguistically-motivated preprocessor module \[12, 6\] in order to perform 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. Classical approaches, such as stemming, rarely manage these phenomena, resulting in erroneous simplifications during the conflation process.

The output generated by our preprocessor is then taken as input by our tagger-lemmatizer, MrTagoo \[9\], although any high-performance part-of-speech tagger could be used instead. MrTagoo is based on a second order Hidden Markov Model (HMM), whose elements and procedures for the estimation of parameters are based on Brant’s work \[7\], and also incorporates certain capabilities which led to its use in our system. Such capabilities include a very efficient structure for storage and search —based on finite-state automata \[11\]—, management of unknown words, the possibility of integrating external dictionaries in the probabilistic frame defined by the HMM \[13\], and the possibility of managing ambiguous segmentations \[10\].

Nevertheless, these kind of tools are very sensitive to spelling errors, as, for example, in the case of sentences written completely in uppercase —e.g., news headlines and subsection headings—, which cannot be correctly managed by the preprocessor and tagger modules. For this reason, the initial output of the tagger is processed by an *uppercase-to-lowercase* module \[23\] in order to process uppercase sentences, converting them to lowercase and restoring the diacritical marks when necessary.

Once text has been tagged, the lemmas of the content words (nouns, verbs and adjectives) are extracted to be indexed. In this way we solve the problems derived from inflection in Spanish. With regard to computational cost, the running cost