Experiments with a Chunker and Lucene

Gil FRANCOPOULO
Tagmatica
www.tagmatica.com

Abstract:
The present paper describes the way we participated to the French track of CLEF-2002.
We used a morphological analyser and a syntactic chunker in order to desambiguate words and filter certain categories of words. Then we built a global index with the Lucene Indexor. Concerning the search process, we wrote boolean queries and evaluated them by the means of the Lucene query Evaluator.

Introduction
The present paper describes the way we participated to the French track of CLEF-2002 during 3 weeks in Spring 2002.
The corpus was composed of two parts. The first corpus was composed of one year production of a news agency whose name is « Agence télégraphique Suisse » and the second corpus was one year of the newpapers « Le Monde ».
It was asked to evaluate 50 queries on an index built from the two corpora.

Indexation
We took two decisions:

First, we decided to improve our « picking up » module that corrects written mistakes. Let's note that these mistakes are numerous in the news texts. The goal was to avoid silence due to the fact that a miswritten word cannot be reached and so cannot propose its text as a candidate during search.

The second decision was to apply a natural language process on the input corpora during indexation.
We has the following reasons to do that:
1) we wanted to recognize compound words as really compound words in order to :
   a) avoid noise due to false interpretation of components: a « pomme de terre » is not a « pomme ».
   b) thanks to the fact that we have a lot of compound words recorded in the lexicon, we can identify words are interesting to index and so, are interesting to identify the document where they appear.
2) we wanted to filter certain grammatical categories, for instance, we wanted to avoid indexation of empty words and adverbs.
3) we wanted to insert inside the index only the lemmatized forms et not the full forms in order to group the various occurrences of the same lemmatized form, and compute a weight for the whole occurrences of the various full forms. This criteria holds for simple and compounds words.
4) we wanted to desambiguate certain difficult (and frequent) French words like « tu » as « Pronoun » vs « Past participle of the verb taire ».
5) we needed to use local grammars in order to recognize dates, times, numbers etc. and the morphological analyzer already had these algorithms.

In other words, we needed to parse the whole sentence.

We proceded as follows:
- sentence segmentation,
- morphological analyses for simple and compound words,
- if the word is unkown, a « picking up » is tried. A rapid visual control has been made on this process. We

verified the unkown words that begin by a lower case and appear more that 5 times and the unknown words that begin by an upper case and appear more that 50 times. The control shows that most of the frequent mistakes were corrected.

- a syntaxic and partial parsing is applied by the means of a chunker (see www.tagmatica.com). We dont use the syntaxic informations labeled by the chunker, we just use the word level desamguisation.
- the lemmatized form is given to the Lucene Indexor (see jakarta.apache.org/lucene).

Search

We translated manually the topics into boolean expressions. Due to the fact that we indexed only lemmatized forms, we expressed the words according to their lemmatized form.
We did not use the title tag. We used only the descriptive and narrative tags.

Conclusion

We don't know exactly how our results can be compared with the results of the other participants.
We indexed the whole ATS corpus but we did not have enough time to index the whole « Le Monde » corpus. We indexed only 70% of the corpus (we had a hardware problem with the machine).
That means that we certainly have a lot of silence compared with the other results. Concerning noise, we probably are not very noisy.