# Integrating New Languages in a Multilingual Search System Based on a Deep Linguistic Analysis

Romaric Besançon, Olivier Ferret, and Christian Fluhr

CEA-LIST, LIC2M (Multilingual Multimedia Knowledge Engineering Laboratory),
B.P.6 - F92265 Fontenay-aux-Roses Cedex, France
{romaric.besancon, olivier.ferret, christian.fluhr}@cea.fr

**Abstract.** The LIC2M has designed a cross-lingual search engine based on a deep linguistic analysis of documents and queries that works on French, English, Spanish, German, Arabic and Chinese. For our participation in the CLEF 2004 campaign, we tested the integration in our system of Russian and Finnish, based on a simplified processing. The results we obtained are not good on the new languages introduced, which shows that our system strongly depends on a correct linguistic analysis of the documents. However, integrating more processing steps in the simplified analysis of new languages so that the results of this analysis are more comparable with the results of the complete linguistic analysis seems to be a good direction for improvements.

## 1 Introduction

The cross-language retrieval system developed at the LIC2M is based on a deep linguistic analysis of both documents and queries. It is currently designed to work on French, English, Spanish, German, Arabic and Chinese. Rather than testing our system on various bilingual tasks on the languages for which we have linguistic resources and processing available, we decided to test, in our CLEF 2004 participation, the possibility of a simple integration, in a limited time, of two new languages: Russian and Finnish. The time factor forced us to use simple strategies for these new languages and try to merge the results obtained with these strategies with the results obtained with the current system.

In section 2, we present the LIC2M multilingual retrieval system: the document and query processing are described, as well as the strategies used for bilingual searches and the merge of the results. We present and discuss in section 3 the results obtained on the different target languages.

## 2 Multilingual Information Retrieval

The LIC2M cross-language retrieval system is a weighted boolean search engine based on a linguistic analysis of the query and the documents. This system has

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 83–89, 2005.

already been used in the small multilingual task of the CLEF 2003 campaign [1]. We present in this section its basic principles.

### 2.1 Document Processing

The documents are processed to extract informative linguistic elements from their text parts. The processing includes part-of-speech tagging and lemmatization of the words and the extraction of compounds and named entities. After part-of-speech tagging, only content-bearing words (nouns, verbs and adjectives) are kept as informative elements of the documents and stored in indexes. This linguistic processing requires the definition of a set of resources for each language:

- a full form dictionary, containing for each word form its possible part-of-speech tags and linguistic features (gender, number, etc);
- a tagged corpus, from which a set of trigrams and bigrams of part-of-speech categories and their frequencies are learned. These trigrams and bigrams are used for the part-of-speech tagging;
- a set of rules for the shallow parsing of sentences. These rules identify the syntactic relations used to extract compounds from the sentences;
- a set of rules for the identification of named entities. These rules are composed of gazetteers and contextual rules that use specific triggers to identify named entities and their type.

The introduction of Russian and Finnish in the multilingual task raised a difficulty concerning this linguistic processing. For Russian, we used a language dictionary that allowed us to simply associate the words with their possible part-of-speech. We had no time to train a part-of-speech tagger nor to develop sets of rules for syntactic analysis or named entities. The processing of Russian just consisted in keeping all possible normalized forms of nouns, verbs and adjectives, with their categories.

For Finnish, since we did not have a full form dictionary, we used a simple stemmer (Porter Snowball stemmer [2]) and no part-of-speech. We also used the stoplist provided by Jacques Savoy [3] to filter out function words and common words. A basic algorithm for decompounding has also been tested. This algorithm considers every word (with a minimum length $l$) that appears in the Finnish corpus as a base word, and splits every word (with minimum length $2 \times l$) that can be decomposed into several base words (all candidate decompositions are kept).

### 2.2 Query Processing

The processing of queries is automatic. Each query is first processed through the linguistic analyzer corresponding to the query language. We kept for this analysis either the three fields of the topic (*title* (T), *description* (D) and *narrative* (N)) or only the two first (T+D).

When using the narrative field in the query processing, a stoplist containing meta-words is used to filter out non-relevant words (words used in the narrative to describe what are relevant documents, such as : “document”, “relevant” etc.).

These meta-words stoplists were built on the basis of a manual analysis of the CLEF 2002 topics.

The result of this analysis is a list of linguistic elements that we call the *concepts* of the query. Each concept is reformulated into a set of *search terms* in the language of the considered index, either using bilingual dictionaries or, in the case of monolingual search, using monolingual reformulation dictionaries (adding synonyms and related words) and/or a topical expansion.

This topical expansion is performed by the same method as the one described in [1]. This method relies on the detection, in a large network of lexical cooccurrences built from a corpus, of the strongly connected components that include the words of the query. The detection is performed in an iterative way by a kind of flow simulation algorithm : a flow starts from the words of the query and is propagated towards their neighbors in the network of lexical cooccurrences to select the words that are the most strongly connected to them. The flow then comes back towards the words of the query to discard those that are not directly linked to the global topic of the query. Finally, it is sent again from the words of the query to select the final expansion words, i.e. the words of the network that are part of the components delimited by the flow and that are not already part of the query.

For translation, we had bilingual dictionaries for French-English and English-Russian pairs. The dictionary we used for the reformulation into Finnish language is the FreeLang bilingual English-Finnish dictionary [4]. Other translations (French-Russian, French-Finnish) were performed through a multi-step translation, using English as a pivot language.

### 2.3 Searching and Merging Strategy

During the query processing, the original topic is associated with four different sets of search terms, one for each target language. Each search term set is used as an independent query against the index of the corresponding language. $N$ documents are retrieved for each language. The $4 \times N$ retrieved documents from the four corpora are then merged and sorted by their relevance to the topic. Only the first 1000 are kept.

For each language, our system retrieves, for each search term, the documents containing the term (until $N$ documents are retrieved). A *concept profile* is associated with each document, each component of which indicates the presence or absence of a query concept in the document (a concept is present in a document if one of its reformulated search terms is present). Retrieved documents sharing the same concept profile are clustered together. This clustering allows a straightforward merging strategy that takes into account the original query concepts and the way they have been reformulated: since the concepts are in the original query language, the concept profiles associated with the clusters formed for different target languages are comparable, and the clusters having the same profile are simply merged.

To compute the relevance weight of each cluster, we first compute a cross-lingual pseudo-*idf* weight of each concept, using only the corpus composed of the $4 \times N$ documents kept as the result of the search. This weight is computed by the formula $idf(c) = \log \frac{4 \times N}{df(c)}$, where $df(c)$ is the number of documents containing the concept $c$. The weight associated with a cluster is then the sum of the weights of the concepts present in its concept profile.

The clusters are then sorted by their weights: all documents in a cluster are given the weight of the cluster (the documents are not sorted inside the clusters). The list of the first 1000 documents from the best clusters is then built and used for the evaluation.

## 3 Results

We tested the system for English and French topics, using $N$=1000 documents retrieved for each language. The result tables present the mean average precision and the number of relevant documents found for each language of the corpus. The average precision for each language is computed only on the subpart of the multilingual search corresponding to the considered language and only on the queries that actually have relevant documents for this language.

### 3.1 Multilingual Results

Table 1 presents the results obtained with English topics using either T+D fields for the analysis of the query or T+D+N, and the results obtained with French topics, using T+D+N fields for the analysis of the query and topical expansion.

**Table 1.** Average precision (avg_p) and number of relevant documents retrieved (relret) for all target languages, using English and French topics

| eng T+D | all | eng | fin | fre | rus |
|---|---|---|---|---|---|
| avg_p | 0.128 | 0.355 | 0.0133 | 0.183 | 0.054 |
| relret | 736 (40.3%) | 235 (62.7%) | 54 (13.5%) | 405 (44.3%) | 42 (34.1%) |
| eng T+D+N | all | eng | fin | fre | rus |
| avg_p | 0.136 | 0.351 | 0.0304 | 0.182 | 0.067 |
| relret | 777 (42.6%) | 240 (64%) | 77 (20%) | 424 (46.3%) | 36 (29.3%) |
| fre T+D+N | all | eng | fin | fre | rus |
| avg_p | 0.126 | 0.18 | 0.0099 | 0.27 | 0.0301 |
| relret | 753 (41.2%) | 157 (41.9%) | 18 (4.5%) | 542 (59.2%) | 36 (29.3%) |

Clearly, our system is weak for Russian and Finnish, the two languages where we did not have a complete linguistic processing and backup solutions were adopted. These solutions are not sufficient to get reasonable results because with its present configuration, our system requires a robust linguistic analysis of the target languages. In particular, the bilingual dictionaries we used for translation

are based on lemmas and parts-of-speech. We should integrate in our system some default processing for the different steps of linguistic processing that would not require the complete definition of linguistic resources but would rely on basic schemas and training data. This would allow to better integrate new languages in the existing design of our system. Another possible improvement is to enrich the reformulation by techniques such as transliteration or approximate matching (for proper names in particular), or use reformulation data automatically learned from aligned corpora.

The results presented in Table 1 also show that our system seems to work better when using all information available in the query (title, description and narrative). The narrative seems to introduce some relevant information by giving different formulations of the topic and without adding much noise after the basic filtering of meta-words by a specialized stop-list. A more precise analysis of the results should be performed to also study the effect of the negative formulations in the narrative ("*documents that contain ... are not relevant*").

### 3.2 Improved Multilingual Results on English and French

Table 2 presents the results obtained for English topics, using T+D+N, only on French and English corpus, after some adjustments of the system that appear to improve the results after a quick analysis of the previous results:

- monolingual reformulation introduces too many rare synonyms (or synonyms of too rare senses of the words) that cause non-relevant documents to be retrieved. For the new test, we simply deactivated this monolingual reformulation (in the future, the monolingual reformulation dictionaries will be checked to improve the relevance of added terms).
- the importance of named entities was neglected in the runs we submitted. Giving a special importance to named entities, relatively to the other words of the query, improves the results. For the new test, we set a double weight for named entities, relatively to the other words of the query.
- the value of $N$ (number of documents retrieved for one language) is also important. Indeed, the documents are retrieved until the number of documents $N$ is reached: if this number is too small, all search terms may not be exploited (search terms are used in the decreasing order of their importance in the collection). For the new test, we set this number at 5000. Notice that the improvement obtained by this adjustment is a trick to improve results in this evaluation framework: using a larger number of documents per language actually helps retrieving new documents at the end of the list, but does not change the first documents retrieved. There are chances that this improvement would not be noticed by a user of a real system.

These results show a significant improvement: the average precision for each language is increased by 25% and 90% of the relevant documents are retrieved.

### 3.3 Finnish Decompounding

The basic decompounding algorithm for Finnish has been tested independently on the Finnish corpus using English topics and T+D fields. Table 3 presents

**Table 2.** Average precision (avg_p) and number of relevant documents retrieved (relret) for French and English corpus, using English topics

| eng T+D+N | fre/eng | eng | fre |
|---|---|---|---|
| avg_p | 0.243 | 0.44 | 0.238 |
| relret | 1168 (90.5%) | 362 (96.5%) | 806 (88.1%) |

**Table 3.** Average precision (avg_p) and number of relevant documents retrieved (relret) for Finnish corpus, using English topics, for different values of minimum length $l$ used for decompounding

| eng T+D | $l = 3$ | $l = 4$ | $l = 5$ | no decompounding |
|---|---|---|---|---|
| avg_p | 0.144 | 0.147 | 0.146 | 0.123 |
| relret | 205 (49.6%) | 190 (46%) | 192 (46.5%) | 181 (43.8%) |

the results obtained[1] with a decompounding using different values of minimum length for base words ($l = 3, l = 3, l = 5$).

These results show that basic decompounding on Finnish tends to improve the results (the gain for mean average precision is 19% but the value is still small) and the best value for minimum length seems to be 4 (though the difference is not important using 3 or 5).

This basic decompounding process is a first step in the design of a more complete simplified linguistic processing for Finnish that could be more compatible with our search system. For instance, we should try to match compounds obtained with this Finnish decompounding algorithm with compounds obtained by a complete syntactic analysis in French and English.

## 4 Conclusion

These experiments in the multilingual track of CLEF 2004 show some improved results of our system, relatively to the CLEF 2003 campaign, on French and English corpora. On the other hand, the poor results obtained for Russian and Finnish show that the introduction of new languages in our system with simplified linguistic processing or stemming/stoplist approaches does not perform well. This integration should be made easier by defining robust default processing for some steps of linguistic analysis so that the results of the simplified processing can be more comparable with the results of the linguistic analysis: the integration of a simple decompounding algorithm for Finnish is a first step in this direction and shows a small improvement of the results. Another direction

[1] These results are not directly comparable with the previous results since they are true bilingual results (not part of multilingual results), and have been obtained with a different version of the search system (that includes a different linguistic analysis for the English queries).

would be to allow the search system to take as input the result of a completely different approach for new languages (for instance, a simple linguistic analysis combined with a reformulation based on statistical translation lexicons learned from aligned corpora). In this case, we would have to tackle the difficulty of merging the results obtained with different processing. Finally, we would also be interested in testing another kind of query expansion based on word senses that are automatically derived from a corpus. We hope that such a resource is more suitable for query expansion than a lexical network such as WordNet that was mainly built by hand.

## References

1. Besançon, R., de Chalendar, G., Ferret, O., Fluhr, C., Mesnard, O., Naets, H.: The LIC2M's CLEF 2003 System. In: Working Notes for the CLEF 2003 Workshop, Trondheim, Norway (2003)
2. Porter, M.: Finnish Snowball Stemmer. http://snowball.tartarus.org/finnish/stemmer.html (2002)
3. Savoy, J.: A Stopword List for Finnish. (http://www.unine.ch/info/clef/)
4. Hämäläinen, K., Kivirinta, T.: Freelang Finnish-English Dictionary. (http://www.kasvua.org/ kphamala/dict.html)