LIC2M experiments at CLEF 2004

Romaric Besançon, Olivier Ferret, Christian Fluhr
CEA-LIST/LIC2M
{*romaric.besancon,olivier.ferret,christian.fluhr*}*@cea.fr*

Abstract

For its second participation in the CLEF campaign, the LIC2M participated in the multilingual task. Our challenge for this participation was to improve the results obtained for French and English and integrate two new languages in the system, Russian and Finnish. Our results are not good on Russian and Finnish, which shows that our system strongly depends on a correct linguistic analysis on the documents.

1 Introduction

The goal of the CLEF multilingual task is to retrieve, in a single list, relevant documents from a multilingual collection. The collection of the CLEF 2004 campaign is composed of documents in English, French, Finnish and Russian.

The cross-language retrieval system developed at the LIC2M is 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 the CLEF 2004 participation, the possibility of a simple integration, in a limited time, of two new languages: Russian and Finnish.

We present in section 2 the multilingual retrieval system we used: the document and query processing and the strategies used for bilingual searches and the merge of the results. We present in section 3 the results obtained for the submitted runs and some improved results obtained on English and French corpora after some simple tuning of our system.

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 been used in the small multilingual task of the previous CLEF 2003 campaign [BdCF$^+$03].

2.1 Document processing

The documents are processed to extract informative linguistic elements from the text parts. The processing includes a part-of-speech tagging of the words, their lemmatization and the extraction of compounds and named entities. 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 set of trigrams and bigrams of part-of-speech categories that are used for part-of-speech tagging (these trigrams and bigrams are learned from a corpus);
- a set of rules for the shallow parsing of sentences. This parsing identifies syntactic relations that are used to extract compounds from the sentences.

- a set of rules for the identification of named entities: these rules are composed of gazetteers and of some contextual rules that uses special 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 has then been quite straightforward since we only used the words and their categories.

For Finnish, since we did not have a full form dictionary, we used a simple stemmer (Porter Snowball stemmer [Por02]) and no part-of-speech. We also apply the stoplist provided by Jacques Savoy [Sav].

2.2 Query processing

All query processing is automatic. Each query is first processed through the linguistic analyzer corresponding to the query language. For two out of the three submitted runs (see section 3), the three fields of the query, *title* (T), *description* (D) and *narrative* (N) were kept for this analysis. For the third run, only *title* and *description* were taken.

When using the narrative field in the query processing, a stoplist containing meta-words was 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 CLEF 2002 topics, from a first selection using frequency information, and revised manually.

The result is a query composed of a list of the linguistic elements extracted from the analysis, possibly filtered by the meta-words stoplists. These elements are called the *concepts* of the query. Each concept is reformulated into *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, based on a network of lexical cooccurrences, as described in [BdCF$^+$03].

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

2.3 Search and Merge Strategy

The search and merging techniques are the same as the ones used in previous CLEF 2003 campaign and are explained in details in [BdCF$^+$03]. They are briefly described in this section.

The original topic is associated, during the query processing, to four different sets of search terms, one for each 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 (in the submitted runs, we took $N = 1000$).

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 term 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 submitted three runs to the multilingual task, described in Table 1. The first two use English topics (one using the title, description and narrative fields, the other using only title and description fields), the third one uses French topics (using title, description and narrative fields for the query processing and topical expansion of the query).

	query language	query fields	query expansion
lic2men1	English	T+D	dictionary reformulation
lic2men2	English	T+D+N	dictionary reformulation
lic2mfr1	French	T+D+N	topical expansion + dictionary reformulation

Table 1: Characteristics of the submitted runs

Figure 1 shows the precision-recall graphs for the three runs we submitted. The global performance of the system is comparable for the three runs.

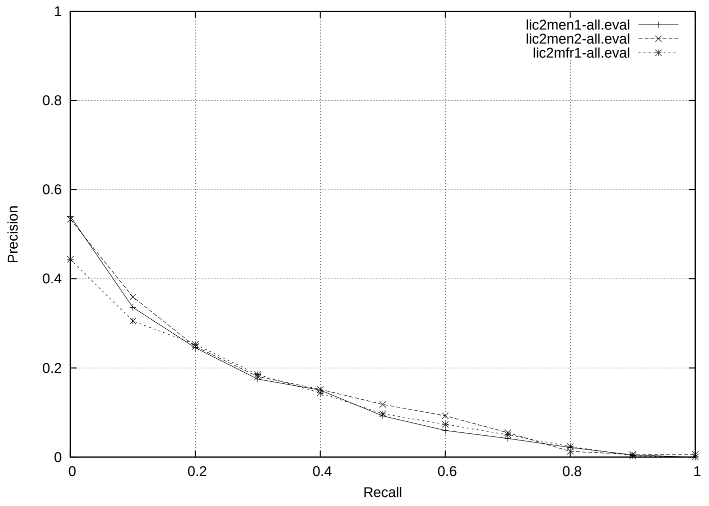

Figure 1: Results of the 3 runs

Table 2 details, for each run, the results obtained on each language independently, given the average precision and the number of relevant documents found. The average precision for each language is calculated only on the queries that actually have relevant documents for this language.

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

lic2men1	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%)

lic2men2	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%)

lic2mfr1	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%)

Table 2: Average precision and number of relevant documents found for each language

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 relies on basic schemas and training data. This would allow to better integrate new languages in the existing design of our system[1]. 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 2 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*").

For French and English, the results are better than for Russian and Finnish but are not as good as we could expect. A first analysis suggests several possible adjustments, that have been tested in a new run:

- 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 other words, improves the results. For the new test, we set a double weight for named entities, relatively to other words.
- 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. For the new test, we set this number at 5000.

With these three changes in the system configuration, the results obtained (for English topics, using the T+D+N fields, and only on French and English corpora) are given in Table 3, and show a significant improvement: 90% of the relevant documents are retrieved. Some tuning of our system remains to be done to improve the ordering of the documents.

4 Conclusion

These experiments in the multilingual track of CLEF 2004 show some improved results of our system, relatively to last year, on French and English corpora. On the other hand, the poor

[1] Notice that this would not solve problems specific to certain languages such as the decompounding of Finnish words.

	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, number of relevant documents found for each language

results obtained for Russian and Finnish show that the introduction of new languages in our system with simplified linguistic processing or stemming/stoplist approaches do not perform well. This integration should be made easier either by making the system more flexible (defining for instance robust default processing for some steps of linguistic analysis) or by allowing the search system to take as input the result of a completely different approach for new languages (for instance, 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 processings. Further experiments in these directions will be undertaken.

References

[BdCF$^+$03] Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard, and Hubert Naets. The LIC2M's CLEF 2003 system. In *Working Notes for the CLEF 2003 Workshop*, Trondheim, Norway, 21-22 August 2003.

[HK] Kimmo Hämäläinen and Toivo Kivirinta. Freelang finnish-english dictionary. http://www.kasvua.org/ kphamala/dict.html.

[Por02] Martin Porter. Finnish snowball stemmer. http://snowball.tartarus.org/finnish/stemmer.html, September 2002.

[Sav] Jacques Savoy. A stopword list for finnish. http://www.unine.ch/info/clef/.