EMIR at the CLIR track of TREC7

Frédérique Bisson, Jérôme Charron, Christian Fluhr, Dominique Schmit

Commissariat à l'Energie Atomique
DIST
CEA/Saclay
91191 Gif/Yvette cedex
France

E_mail : fluhr@tabarly.saclay.cea.fr

1 Introduction :

EMIR (European Multilingual Information retrieval) was a European ESPRIT project whose aim was to demonstrate the feasibility of a crosslingual interrogation based on the use of bilingual dictionaries. The project lasted from November 90 to April 94. A part of the results are included into a commercial product " SPIRIT " released by the T.GID Company in France.

2 Basic Principles of EMIR :

Principle of relevance ranking :

The system can be considered as a weighted boolean system. That means that the result of an interrogation is a partition of the database into classes of intersection, each of them being identified by the best boolean query that can retrieve the documents of this class.

Intersections are weighted according to the weight of each single word or compound that is in the intersection. The weight of each word is related to its discrimination power. That means that words pertaining to few documents have higher weights than words that are in a large number of documents.

Linguistic analysis :

The documents and the queries are processed by a linguistic analysis that normalizes words (synonyms are represented by the same character string and some homographs are solved and represented by different character strings). Compounds are recognized and normalized. To each normalized word or compound is associated a part of speech.

Reformulation :

To increase the relevance, a query expansion based on monolingual or bilingual reformulation rules is used. Each query word or compound can infer, according to its part of speech, synonyms or words derivated from the same root or translations into an other language.

The intersections are evaluated on the base of the original query words, that means that any of the inferred words from one query word can represent it.

3 Lessons from the previous TREC :

Our results in the TREC6 for the crosslingual track, can be considered not very good for the monolingual interrogation but the difference between mono and crosslingual interrogation was small.

So we have done a study to identify the causes of the level of monolingual results. In fact if we solve this problem we can also increase our results in crosslingual interrogation.

Several problems were identified :

Definition of relevance :

We have not the same definition of relevance than the one used in the crosslingual track. That means that some documents we consider relevant are considered in the "qrels" as non relevant. Probably the reason is that our system is tuned to access information and not to access documents. That means that if only one line in a 200 page document is relevant, we consider the document as relevant.

Ranking of classes :

The ranking of classes is not optimal. One of the main cause is that some compounds which are not important for the query but have a high weight because they are in few documents, can give a better weight to a non relevant intersection and push the good ones to the bottom of the list.

Incompleteness and inconsistency :

An other less important point, but that must be fixed, is the incompleteness and inconsistency of the various dictionaries and reformulation rules. They have a bad effect on both monolingual and bilingual reformulation.

Lack of a multi step reformulation :

In some cases bilingual reformulation must be followed by a monolingual one in the target language. It is especially the case when there is a change of part of speech in the translation of a word. For example an adjective can be translated as a noun.

Ex:

debt for Poland -> dette polonaise (Query 33)
génie génétique -> genetically engineered (Query 38)

4 The situation for TREC7 :

For TREC7 like for TREC 6, runs have been done using only the “desc” part of the topics. Misprints detected by the linguistic processing have been removed. A monolingual reformulation has been performed on the part of the database which is in the same language than the query. A multilingual reformulation without target language monolingual reformulation has been performed when the query and database language are different.

Concerning the definition of relevance, we don't want to change our definition of relevance because our users are happy with this definition.

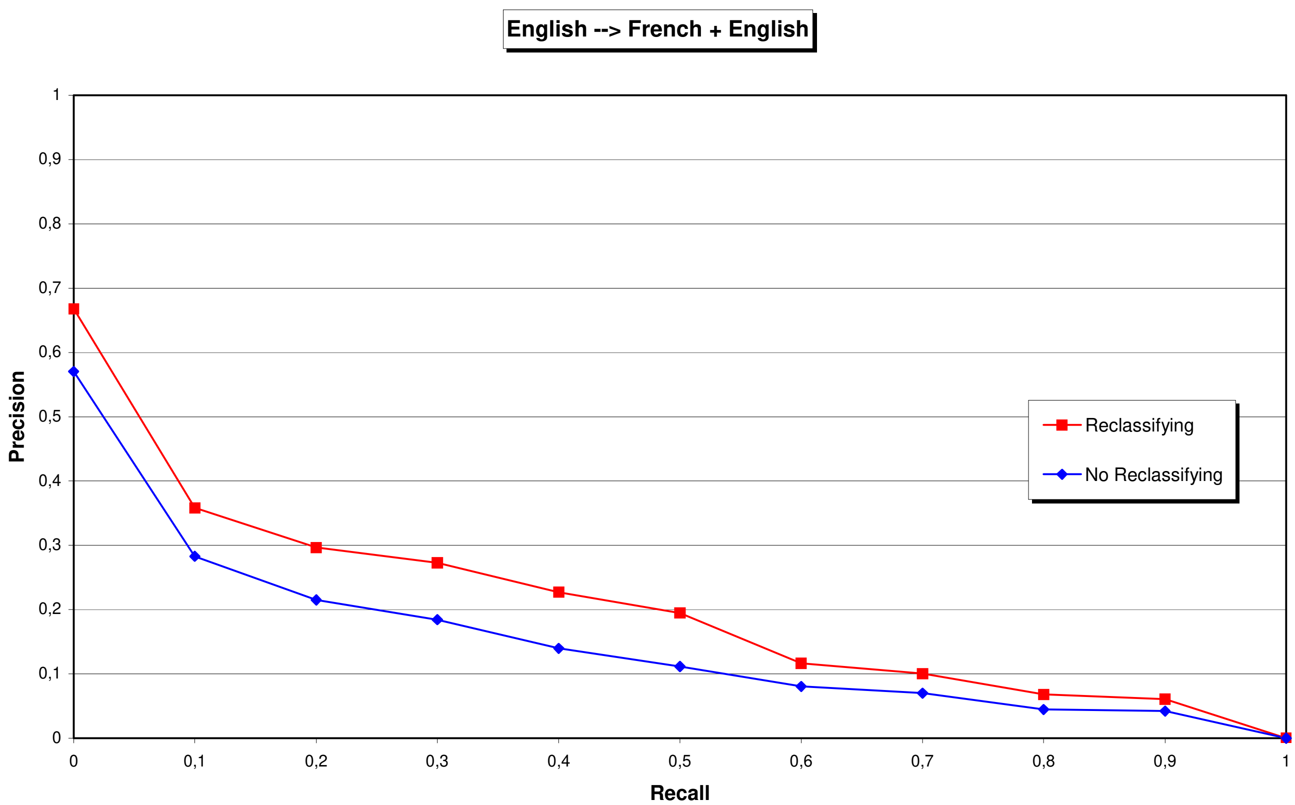
Concerning the ranking of classes, we think that there is two ways of improving the system.

The first one is by applying a better program for compound recognition and we have already one which is not implemented into the commercial product. The one currently used can give wrong compounds. This kind of wrong compounds are typically compounds that can disturb the document ranking.

Ex:
"oil from point " in query 27 (… method of delivering oil from point of origin to the shipping points …)

Another reason of bad ranking is due to the fact that, in long queries, it is necessary to combine the weights computed from the database with weights measuring the importance of the words for the user. For example in the query about Lötschberg, if this proper noun is not in the intersection between a document and the query, the document is surely not relevant.

To experiment this hypothesis we have done two runs, one using the only weights computed from the database like last year and an other run where intersections that do not contain a compulsory word for the query are automatically pushed after the last intersection containing the compulsory word.



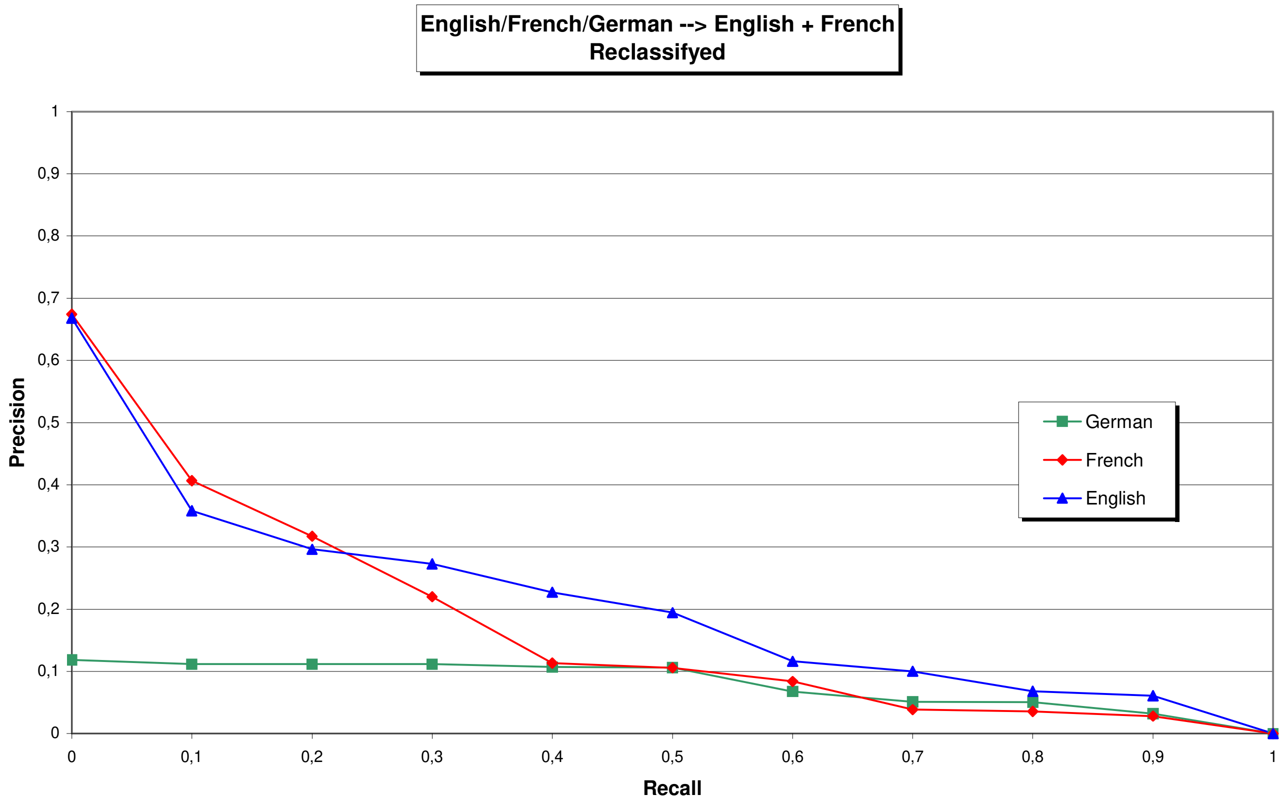
About the incompleteness, we have done a study of the vocabulary of the databases (both English and French) : unknown words have been added to the dictionary and in some cases we have introduced a compound into the automatic expression dictionary. Lacking translations have been added.

About inconsistency, it is a very important problem when large linguistic data is to be managed. At this time we manage monolingual dictionaries (single words and idiomatic expressions) for 5 languages, monolingual reformulation rules for 5 languages, bilingual reformulation rules for 9 couples of languages.

The monolingual dictionaries contain the relation between each word form and a normalized form that is used for retrieval. The normalized form is taken out of the list of equivalent forms. This includes flexion from the same lemma but also absolute synonyms like orthographical variations like “colour” - “color”. Of course word forms inside the reformulation rules must be consistent with the choice of the normalized form.

Maintaining a full consistency manually is impossible. We are developing an architecture to ensure a consistency validation. At this time the development of this application is not finished and we are sure that in our runs for TREC7 some lack of intersection are due to this bad consistency.

This is especially visible in the results of the run with German queries because we have received from our German partner (TEXTEC in Saarbrücken) the bilingual dictionaries few days before the deadline and it was not possible to ensure consistency between the monolingual German dictionary and the bilingual reformulation one.



At last, as we have used the standard SPIRIT server, it was not possible to use the multistep reformulation tested in a prototype that is necessary to make a monolingual reformulation after the bilingual one. We expect that the functionalities of the prototype will be introduce before next TREC into the standard server.

5 Software architecture :

At the occasion of TREC 7 we have experimented in a larger way the multilingual, multibase architecture we intend to put into service at the end of this year. The French part of the database has been generated into 3 different databases. In that way it was possible to generate these 3 parts in parallel. The same was done for the English database that was split into 3 parts and generated in 3 different databases in parallel.

The interrogation is done through a web viewer. Databases in the same language as the query are interrogated using monolingual reformulation. Databases in a different language are interrogated using bilingual reformulation. All interrogations are done in parallel.

The interface between the internet viewer and a standard SPIRIT server executes the merging of results and the computation of a global weighting.

This merging is highly facilitated by the fact that the intersection is characterized by the words from the original query.

6 Automatic interrogation :

TREC 7 was also the occasion to test the system for automatic interrogation of information retrieval systems on the web. This system, which is named BeFor (Beyond Forms), has been developed by Jérôme Charron during his PhD. This system uses an XML description of the interrogation screen, an XML description of queries that can be more complicated than the one from TREC and can merge factual information and full text search, an XML description of how to use the result of the search.

In our case we have used this system to make an automatic run of TREC queries to the Web interrogation interface we have used to test TREC.

This system can manage at the same time several sets of queries and several interrogation applications. The extraction of results is converted into a TRECEVAL format.

It has been chosen by the organizer of AMARYLLIS to make an automatic run through internet directly by the organizer without any possibility of intervention by the author of the tested IR systems.

7 References :

About reformulation in full text IRS, F. Debili, C. Fluhr, P Radasoa, Conference RIAO 88, MIT Cambridge, march 1988, Modified version has been published in "Information processing and management" Vol. 25, N° 6 1989, pp 647-657.

EMIR Final report, Christian Fluhr, Patrick Mordini, André Moulin, Erwin Stegentritt, ESPRIT project 5312, DG III, Commission of the European Union, october 1994

Multilingual database and crosslingual interrogation in a real internet application, C. Fluhr, D. Schmit, F. Elkateb, K. Gurtner, workshop "Cross-language Text and Speech retrieval" in "AAAI 1997 Spring Symposium Series", 24-26 march 1997, Stanford University, California.

SPIRIT-W3, A distributed crosslingual indexing and retrieval engine, C. Fluhr, D. Schmit, Ph. Ortet, F. Elkateb, K. Gurtner, INET'97, Kuala Lumpur, June 1997.

-

EMIR at the crosslingual track of TREC-6, F. Elkateb and C. Fluhr, TREC-6 Conference, 19-21 November 1997, Gaitherburg, Maryland

Cross-language information retrieval, Grefenstette, G. and alii., Boston: Kluwer Academic Publishers, 1998.