Multilingual database and crosslingual interrogation in a real internet application

Architecture and problems of implementation

Christian Fluhr, Dominique Schmit, Faïza Elkateb, Philippe Ortet, Karine Gurtner

Commissariat à l'Energie Atomique
Direction de l'Information Scientifique et Technique
CEA/Saclay
91191 Gif sur Yvette cedex
France
E_mail : fluhr@tabarly.saclay.cea.fr

Abstract

The EMIR European project demonstrated the feasibility of a crosslingual interrogation of fulltext databases using a mono and a bilingual general reformulation of the query.
New developments have been done to accept multilingual databases even if there is more than one language in the same documents.
Problems of implementation of full-scale applications are discussed.

Background

Our crosslingual text retrieval technology is based on the results of the EMIR (Fluhr 1994) (European Multilingual Information Retrieval) project in the framework of the ESPRIT European Program . The project was done on three languages : English, French and German, with partners from France, Germany and Belgium. The project was performed from October 90 to April 94.

The project had as technological support an already existing multimonolingual (French and English) text retrieval software : SPIRIT (Syntactic and Probabilistic Indexing and Retrieval of Information in Texts). SPIRIT has been marketed since 1980 on IBM mainframe, since 1985 on various platforms (from PC to mainframes) and from 1993 in a client-server architecture.

At this time a part of EMIR results are introduced in the SPIRIT system. Applications are especially for the French-English couple. A new language has been added : Russian and the Dutch language is on the way.

Main principles of the approach of crosslingual interrogation

The EMIR's approach is based on the use of a general transfer dictionary as a set of reformulation rules. That means that all possible translations are proposed as possible key for retrieving the documents.

This approach is opposite to the one consisting of a translation of the query followed by a monolingual interrogation.

There are 3 differences :

- the reformulation is based on the translation of concepts in the query and need not to build up a syntactically and semantically correct translated query.

- the translation of concepts is used to obtain documents, if the word inferred in the target language is not the right translation but if a relevant documents is obtained we can consider that the system works properly (for example : inference of an hyperonym or a word of a different part of speech).

- The translation and retrieval process are mixed so that the fulltext database is used as a semantic knowledge for solving translation ambiguities which are the main problem to solve in this approach of domain independent reformulation. One of the main result of this approach is that if an answer to the query exists in

the database, the system, in most cases, can select the right translation by looking in most relevant documents.

Principle of the method

We suppose that the database is monolingual, we will discuss in the following paragraphs the problems specific to multilingual databases.

Database processing

The database is processed by the morphosyntactic parser. The results are normalized single words and compounds. Normalization is mainly based on a lemmatization but general synonymies can be taken into account. For example « logiciel » and « software » in French are normalized by « logiciel », « harbour » and « harbor » in English are normalized by « harbor ».

As single words we assume really single words and idiomatic expressions like « monkey wrench » in GB English or « clé anglaise » in French. Compounds are words in dependency relations like « multilingual database ». For each normalized word or compound a semantic weight is computed according to the information it brings to choose the relevent documents.

Query processing

The query is processed by the same morphosyntactic parsing than for the database. Normalized words and compounds are produced with their part of speech.

For each of these units, we try to infer all possible translations that agree with the part of speech. For example « light » adjective is translated by « léger » adjective in French but « light » noun is translated by « lumière » noun.

Compounds can be translated globally or word for word. In this last case the word order is rearranged to fit the result of the target language normalization of compounds. All compounds that cannot be translated word for word but it is not necessary to consider them as idiomatic expression. A compound like seat belt is really a belt on a seat and in French « ceinture de sécurité » is a belt for security.

Generally, especially for single words, there is a lot of translations.

Example : « talon » (French)--à(English) « heel », « crust », « spur », « stub », « conterfoil », « talon »

At this level the results of multilingual inference is filtered by the database lexicon and a lot of translations that are incompatible with the domain are eliminated.

The filtering by the database lexicon is not sufficient to eliminate all wrong translations. So it is possible to take the translations contained in the most relevant documents (that means the ones that contain the maximum of the query words, especially the ones where words have the same dependency relations than in the query).

It is necessary before performing this optimization to be quite sure that the « best » documents are relevant, that is to say that they contain a sufficient number of the most weighty words.

If it is decided that the most relevant documents are really relevant ones, a feed back can be done on the transfer process. In a second step, only words compatible with the most relevant documents are proposed.

This process is very strong to increase relevance but it has a bad effect on the recall because it can eliminate synonyms of the chosen words that are only in less relevant documents. So it is useful to follow this feed back by a monolingual reformulation in the target language. We are in the same situation that a well formed query directly in the target language or a well translated query that necessitates a monolingual reformulation to have a good recall.

Example of translation and filtering

Query : « spectroscopie de masse par temps de vol » on a base of 655000 titles of reports on Energy

transfer rules :

spectroscopie (Noun) : spectroscopy
masse (Noun) : mass, bulk, ground, sledgehammer
temps (Noun) : stroke, tense, beat, time, weather, days
vol (Noun) : flight, flock, theft

After filtering by the database lexicon :

spectroscopie (Noun) : spectroscopy
masse (Noun) : mass, bulk, ground
temps (Noun) : stroke, beat, time, weather, days
vol (Noun) : flight, theft

After filtering by the best document that contain : « time of flight mass spectroscopy » where the system has recognize 2 compounds « time of flight » and « mass spectroscopy »

spectroscopie (Noun) : spectroscopy
masse (Noun) : mass
temps (Noun) : time
vol (Noun) : flight

that means that the system has dynamically produced that :

temps de vol : time of flight
spectroscopy de masse : mass spectroscopy

The rearrangement of words after word for word translation has been obtained using rules depending from the couple of languages and the syntactic structure of the compounds.

Architecture for multilingual databases on INTERNET

The EMIR project has demonstrate that it is possible to query a monolingual database in a language that is different than the database language. But in our countries where we use our own language and English for scientific documents, the problem is that most of the databases contains documents in 2 or mores languages. Even in some cases, the information about a document is in several languages. For example, in our library catalog, a document can have an original title in English, a translated title in French, keywords in English and French, a summary in English and French.

Generally it is not possible to assume that the information is redundant between the languages. For example, keywords in English and in French are taken from different systems of indexing and they are not translations from each other.

The mixed language databases have not been taken into account by EMIR and nor in the current version of SPIRIT system. That is the reason why, the problem has been solved through a new architecture based on a more standard Web architecture.

For permitting the access of SPIRIT server using standard INTERNET clients (Netscape, MS explorer, tango, etc.), a WWW-SPIRIT interface has been done at the end of 1995.

This interface is being extended to support distributed multibase multilingual databases. That means that the system can manage a logical database composed of several physical ones. The problem of mixed language database is supported by this architecture in the following way :

- The mixed logical database is split into as many physical databases as languages in the logical one..
- the same documents can have parts in several monolingual physical databases
- during the interrogation the query in one language (for example the user's mother tongue) is sent to each physical database composing the logical mixed language one.
- each database server receive the query with an information of what is the query language. The database server knowing the language of his database performs either a monolingual or a bilingual interrogation and send results to the interface.
- the interface must merge the results, for that it computes a weight of each concept of the query on an hypothetical database composed of the retrieved documents from the various language parts. This processing is necessary because the weight of each concepts computed in each part can be very different and cannot represent the weight of the global use of the concepts in the multilingual logical database.
- In SPIRIT the answers are grouped by intersection classes and the classes are sorted according to the weight of concepts in the intersection query-document. The merging of results can produce new classes and suppress others.
- at the end when a user ask to see a document, parts from different languages are to be obtained from various monolingual physical databases. The location of word that is also obtained from various monolingual physical databases are used to highlight the words used to extract the document. This functionality is specially important in the case of crosslingual interrogation to control why a documents has been retrieve.

First applications

At this time our library catalog are put on the web using the first version of the SPIRIT/W3 interface which cannot manage multilingual databases nor multibase interrogation.

Our documents are mainly in French and English. We are obliged to choose only one indexing language, for example French. For this indexing language, all the linguistic support is given (lemmatization), for the other language all words are considered as proper nouns and are not lemmatized.

It is easy to show that to be sure to have a sufficient answer it is necessary to ask a simple query in French, and a query with all possible variants of the words in English.

Thanks to god, main of our users are happy with partial answers and have not seen the problems. But it is not a good idea to expect that this situation will continue.

That is the reason why, we decided to implement in an operational way the prototype of the SPIRIT/W3 multilingual interface. We aim at beginning the service for late march 1997 on our library catalogs, catalogs of reports on energy from the IAEA agency and from ETDE and a catalog of the publications done by personal belonging to our organization. All these applications will be visible from the all internet. Unfortunately, fulltext databases will be reserved for intranet use.

Problems to solve

Going from a feasibility prototype to an operational software is a hard task especially when the problem to solve is crosslingual interrogation.

Such a systems involves a lot of tools and a lot of linguistic data. If somewhere something lacks, the result is strongly perturbed. That is the reason why a very strong quality control is necessary to minimize the possibility of discontinuity between the query character string and the words in an other language to search in the database index.

Control quality on linguistic data

There are many causes of the discontinuity in the linguistic process :

- the word is not in the dictionary of the source language
- the word pertain to an idiomatic expression which is not in the source language dictionary.
- the word is in the source language dictionary but not with all the possible part of speech (example ; processing as a verb but not as a noun)
- the normalized word has no entry in the transfer dictionary
- the word cannot be translated word for word : a translation of compound must be put into the transfer dictionary.
- normalization of words in the source language is different from the normalization of the left part of the transfer dictionary.
- compound are in different parts of the linguistic data considered as simple compound (words in dependency relations) or idiomatic expressions and in SPIRIT their normalization can be different.

This list is not exhaustive.

In a linguistic process which goes through many steps, a lack of coherence can brake the inference line. When managing a huge amount of linguistic data, it is necessary employ many people from different origin and even working in different part of the world.

So it is not possible to trust only in the human behavior. The solution is to implement a control system who can verify all that can be verify automatically and suggests to the human corrections or additions of information.

Acquisition of new compounds in the transfer dictionary

An other need is the fast adding of new words for new databases. It is very easy to know what single word lack in the source language dictionary and in the transfer dictionary. The main problem is what compounds that cannot be translated word for word must be introduced into the transfer dictionary.

Various strategies have been followed to tackle this problem :

- automatic extraction of terminology and treatment by a specialist.
- treatment of bilingual corpuses in the same domain, extraction of terminology in the two lingual version and determination of the compounds that are not translatable word for word. This way seems promising but cannot be followed at this time because of the problems of consistence in the linguistic data mentioned above. At this time too much false detection are obtained to permit a use in exploitation
- processing of multilingual thesaurus in the domain of the database. We have begun to process the INIS (IAEA) and ETDE (OECD) thesaurus. The first one is on atomic energy, the second one is on all kind of energy.

Conclusion

It is difficult at this moment to give general conclusions. Our evaluation done during the EMIR project was that even

with a lack of consistency and without adding any word into the dictionaries and without translation feed back, we had a decreasing of results of about 10 % in comparison with a monolingual interrogation with the same system.

In comparison with the actual situation where databases are indexed in only one of the languages, the new architecture will increase the quality of the answer using only one query in one language. What will be the level of the result in comparison with really good translation in each languages, the evaluation is to be done.

In fact, even if the level of performance is not the best one, according the behavior of our users, (they ask only one question in one language), the results will be strongly ameliorated in comparison with the actual performance of the database.

References

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

Fluhr C., Multilingual Information, Pacific Rim International Conference on Artificial Intelligence (PRICAI), "AI and Large-Scale Information", Nagoya, 14-16 November 1990.

Fluhr C., Radwan Kh., Fulltext databases as lexical semantic knowledge for multilingual interrogation and machine translation, EWAIC'93 Conference, Moscow, 7-9 September 1993.

Fluhr C., Mordini P., Moulin A., Stegentritt E., EMIR Final report, ESPRIT project 5312, DG III, Commission of the European Union, October 1994

Fluhr C., Schmit D., Ortet P., Elkateb F., Gurtner K., Semenova V., Distributed multilingual information retrieval, MULSAIC Workshop, ECAI96 Conference, Budapest, 12-16 August 1996

Gachot D., Lange E., Yang J., Teh SYSTRAN NLP Browser : An Application of Machine Translation Technology in Multilingual Information Retrieval, Cross-Linguistic Information Retrieval Workshop, SIGIR'96, August 18-22, Zurich, Switzerland.

Landauer T. K., Littman M. L., Fully Automatic Cross-Language Document retrieval Using Latent Semantic Indexing, (1990) in Proceedings of the sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and text Research, UW Centre for the New OED and text Research, Waterloo Ontario Canada

Radwan Kh., Foussier F., Fluhr C., Multilingual access to textual databases. RIAO'91 Conference, April 1991, Barcelona.

Radwan Kh., Fluhr C., Textual database lexicon used as a filter to resolve semantic ambiguity, application on multilingual information retrieval, 4th annual symposium on document analysis and information retrieval, Las Vegas, 24-26 April 1995.