Crosslingual Interrogation of Multilingual Catalogs

C. Fluhr, D. Schmit, C. Andrieux, Ph. Ortet, F. Bisson, V. Combet

CEA/DIST
Centre d'Etudes de Saclay
91191 Gif/Yvette Cedex, France

(fluhr,schmit,andrieux,ortet,combet)@cartier.cea.fr

Abstract. In this paper, we describe a crosslingual Information Retrieval System (IRS), which makes the interrogation of multilingual databases possible. Indeed, the CEA needs to be able to process an important amount of multilingual databases and documents, and so we had to adapt the IRS we use, SPIRIT (which relies on linguistical and statistical processing) to this situation. We have thus set up a crosslingual interrogation system based on the indexation of documents containing parts written in different languages and on the bilingual reformulation of the query. The latter tries all the possible translations for every significant word of the query and the documents are used as filters in case of uncertainty or ambiguity. The answers to a query are given in the form of a list of classes of documents ranked according to their relevance. This paper describes the application of these techniques to crosslingual access to catalogs and bibliographic databases.

1 Introduction

In countries where English is not the sole official language, the scientific information managed by public and private bodies is conveyed both in English and in vernacular. Of course, other languages are used, for instance, in our organization, German, Japanese, Russian or Spanish. But this can be considered as a marginal phenomenon in comparison with the use of English and French.

Even for publications issued by our scientists and that constitute a part of our organization's scientific memory, the number of English documents is important.

A large number of our databases contain both English and French documents: book catalogs of our libraries, a catalog of articles from about 3,000 scientific journals, a bibliographic database of international reports on energy and a bibliographic database of the publications issued by our scientists.

In addition to this, it is worth mentioning the databases which are built up from Web pages, generally in several languages, extracted for a technology watch on a particular theme. But this is not the subject of this paper.

For the majority of users who are likely to connect to our Intranet, the information retrieval systems that focus their search upon character strings are difficult to use with a good effectiveness. This is why access to information in our organization is based upon a sophisticated information retrieval system using morphosyntactic parsing to process both documents and queries. A statistical model is also used to weight normalized words. As a result, the answer to a query is a list of documents ranked according to their relevance, grouped in classes characterized by the same concept intersection with the query.

This works well for monolingual databases. In the case of bilingual information, however, as automatic linguistic processing produces normalized words thanks to the dictionaries and syntactic description of a given language (L), a simple application of this technique is not straightforward. Indeed, the part of the database which is in language L is analyzed correctly and thus well indexed. In the remaining part however, the words in the other language are not normalized, except for words that seem to belong to the main language (L). In the last case, the result of normalization is incorrect. For example, if the database is indexed in French, the word “data” will be considered as the word "dater" that means "to date". One can easily imagine what kind of strange inferences can be drawn after an automatic expansion of the query.

Fortunately those cases are not the most frequent ones. Yet, as a whole, search quality has not reached the level that could be expected from a sophisticated natural language information retrieval system.

This is why we took in 1990 the initiative of the first crosslingual information retrieval European project within the framework of ESPRIT: EMIR (European Multilingual Information Retrieval). Part of the results have been introduced into the SPIRIT system, a product marketed by T-GID that we use for all our applications.

The project demonstrated the feasibility of a crosslingual interrogation using a bilingual reformulation that tries all possible translations of each query word. In the project, the database queried was monolingual. Now, it is not the case in real life, and so, since the end of the project until now we have been developing technologies to access mixed language databases and improve the quality of bilingual reformulation.

2 Description of the Data to Be Accessed

We want to limit this presentation to the problem of data retrieval from catalogs and bibliographic databases. Of course, we are also involved in the crosslingual interrogation of fulltext databases, both in the case of project memory (all the texts interesting a consortium within the framework of a scientific project) and in the case of scientific, technological or strategic watch.

Using catalogs to search information on a given subject is a problem much more difficult to tackle than retrieving a known title because of the very little information available (titles, and in some cases, controlled or free keywords, and summaries). On the contrary, fulltext contains a large number of redundancies and thus offers more chances of finding the relevant documents.

We will focus our presentation on two different cases. One is the catalogs of our libraries. The second is a bibliographic database of the publications issued by our scientists.

The catalog items are filled using the OCLC database. Some local information is added. The proportion of French books is 12 %, English books 75 %, Russian books 3.3 %, and German books 2.3 %. The rest is negligible.

The main problem to be tackled is the following : even if our source is the OCLC catalog, the content description is not in only one language. Titles are in the original language (mainly French or English) but there are keywords in a different language than the original one, for instance English keywords in the case of French books. Keywords in French are assigned by the Library of France which chooses them from the Rameau Thesaurus. Of course, English and French keywords are not translations of one another.

Example :

Title : Automating library procedures, a survivor's handbook
Author : Ian Lovecy

English keywords : Libraries, Automation, Library science, Data processing, Library, Administration

French keywords : Bibliotheques, Informatique, TRAITEMENT DE L'INFORMATION

The bibliographic database of CEA scientists' publications contains original titles. The database content is keyed in directly by researchers' scientific departments. They can add the title in the other language (French if the document is in English or English if the document is in French), a summary in French and/or English and free keywords in French and/or English.

When summaries or keywords are in both languages, they are not always translations from one another.

Given the two following facts :
- the bilingual parts of the description do not always result from a translation,
- linguistic processing is not a technology giving 100% of right results,

it is preferable to make use of all the information available for document access. As a result, if the system failed to find a document using French information, the same document could be retrieved using English. That is the reason why we have decided to build up an architecture that can exploit the full information available in each document description so as to compare it with the query, generally expressed in one language. The semantic intersection between the query and documents is reached taking into account both languages.

Referring to both languages at the same time ensures an efficient ranking of mixed language documents.

3 Crosslingual Approach of the EMIR Project

3.1 Short State of the Art

There are three main approaches to crosslingual interrogation of text databases.

The first one is based upon a statistical approach assuming that a number of documents of the databases are in both languages. In some cases documents do not necessarily have to be translations from one another, provided that they are identified as being related to the same topic or event, as it is the case with newspaper articles.

The most well known statistical model is the vector space model of Gerard Salton [2]. In this model the database is represented as a vector space which has as many dimensions as the number of different words in the database. Documents or queries are vectors in this space. A proximity is evaluated between a document and a query by computing the cosine of the angle between the two vectors.

An improved model is the latent semantic indexing one (LSI) [3]. By reducing dimensions to a few hundreds, it achieves a sort of implicit reformulation both on each language and between the languages involved.

The results are good but suppose that translated documents exist.

The second approach is based on the translation either of the queries or of the full database (generally the query) using a machine translation (MT) system. The drawback of such an approach is due to the fact that MT systems provide one translation for each word : if they give the wrong translation (as it is often the case for polysemic words), the query result may be bad.

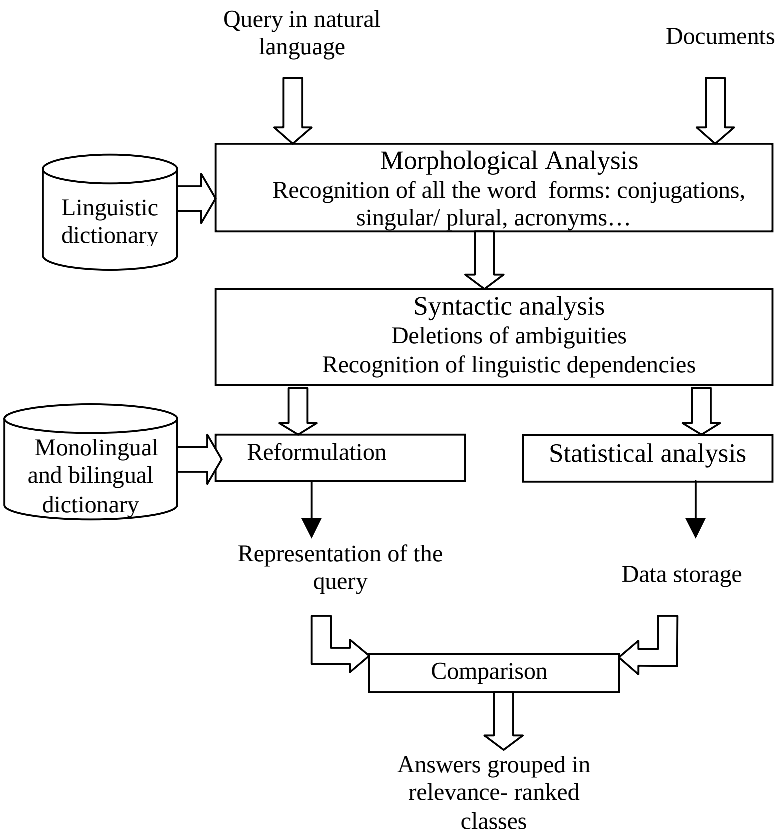
The third approach, which is ours, consists in using a bilingual reformulation that tries any possible translation for each word. The ambiguities are solved if there is an answer to the query by using the relevant documents as a semantic filter to choose the right translation. Of course, this is only the case if the query contains several words.

3.2 SPIRIT's Basic Principles

The EMIR (European Multilingual Information Retrieval) project was based on the technology of an existing product, the SPIRIT System. SPIRIT stands for "Syntactic and Probabilistic Indexing and Retrieval of Information in Texts".

SPIRIT components are the following:

Fig. 1. SPIRIT Components



The linguistic processing consists of a morphosyntactic parsing that assigns a part of speech to each word, recognizes idiomatic expressions using a dictionary, normalizes words through lemmatization, recognizes dependency relations (especially those involved in compounds), identifies general language synonyms (ex : Uranium 235, U235, U 235, U5) and solves some homographic problems (especially those that can be solved by syntactic parsing like "train" verb and noun). At the end, empty words are rejected according to their part of speech (prepositions, articles, punctuation, etc.).

The statistical model is used to give the user a list of documents sorted according to their degree of relevance. The SPIRIT model differs from the vector space model because it assigns a weight to each word of the database according to its discriminating power, but does not assign a weight to each word in each document.

This point is important because the aim of the system is to find relevant information even in documents in which the main content is out of scope of the query.

The SPIRIT system can be considered as a weighted Boolean system. This means that documents are grouped into classes characterized by the same concept intersection with the query. The classes constitute a partition of the database, which implies that each document is in the best class it can be. Each class of intersection is characterized by a boolean query where only "AND" and "AND within a dependancy relation" are used as operators. Of course the second operator is considered as better than the "AND" one. In the following example, documents of the first class can be also accessed by the queries characterizing the following classes. But the document is only put in the best class (that means the first one in the decreasing order of relevance) it can be in.

Example:
Question: management of radioactive wastes

Result:
First class: 166 documents, characterized by "*management of radioactive wastes*"
Second class: 62 documents, characterized by "*management of wastes and radioactive*"
Third class: 4 documents, characterized by "*radioactive wastes and management*"
Fourth class: 44 documents, characterized by "*management of wastes*"
Fifth class: 356 documents, characterized by "*radioactive wastes*" etc.

Reformulation is used to infer from the original query words other words expressing the same concepts. Reformulation can be in the same language (synonyms, hyponyms, etc…) or in a different language (bilingual dictionary).

The comparison tool is used to perform a quick evaluation of all possible intersections between documents and query words (including those inferred through reformulation), and to compute a relevance weight for each document. For information retrieval the class weight only depends on the query/document intersection. The weight of each concept (that means all words inferred from one original query word) depend only on the number of document where it occurs.

This approach differs from a standard query expansion because the intersection is based on the original query words and not on the set of words inferred from the query. That is why we assume it as a concept intersection instead of a word intersection.

3.3 General Principles of Crosslingual Interrogation of SPIRIT Using Bilingual Dictionaries:

Any possible translations are inferred from the original query words. Some of these translations are not in the database and are naturally dropped. For the rest, they are filtered by the database itself as their co-occurrences are taken into account.

In fact, if a document is relevant to the query, the intersection between the concepts expressed by the query words and those expressed in the documents will be broad. In this case the co-occurrence of translations from a maximum of different query words gives the quasi-certainty to get the right translations of the query words.

Example :

If the query contains “ free space ” where “free” is tagged as an adjective and “space” as a noun by the syntactic disambiguation:

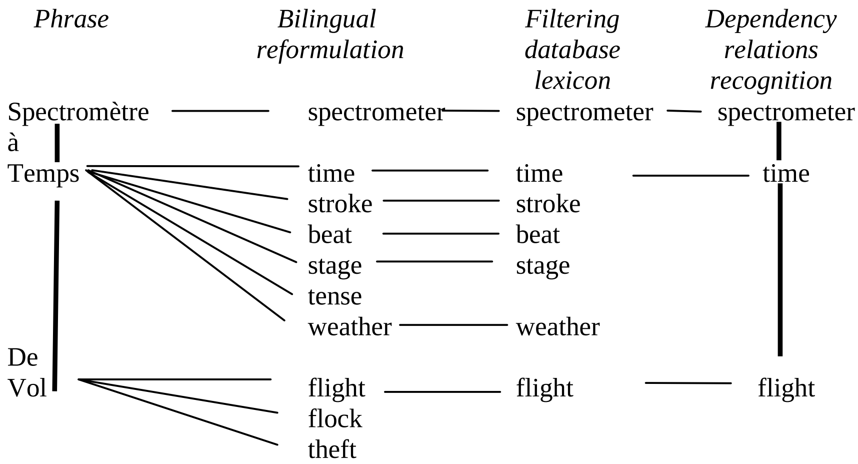
The possible translations of "free" (adjective) into French are: gratuit, large, libre.

The possible translations of "space" (noun) into French are: blanc, espace, place, période.

The only translations of “free space” that can be found in the text database are “espace libre” or “place libre” which are both acceptable translations.

Of course, the dependency relations are taken into account. Indeed, generally a two words compound in a dependency relation can have more than one hundred translation candidates, using word for word translation and a reordering of the translated words - but only few of them can occur in the database with the same dependency relation. More details can be found on the EMIR’s crosslingual interrogation in [4].

Table 1. Example



Vertical links represents dependency relations.

This schema shows the different steps that are used to infer the right translations using the database as a semantic filter. At the first stage, the system infers any possible translations of all single words or compounds (if there is a global reformulation for a compound). In our example there is no reformulation of compounds because they can be translated word to word.

After that, translations that do not appear in the database are eliminated. For example, "theft" and "flock" for "vol", "tense" for "temps".

To find the right translation of compounds translated word to word, it is necessary to use transformational rules to eventually reorder the words in the target language. That means for example that the system try to find either "time of flight" and "flight time" as possible translations of "temps de vol"

The last filtering step is done by examining the translations appearing in the "best" documents that means the document having the greatest intersection with the query. The "best" class of intersection is characterized by the words in dependency relation "spectromètre-temps-vol". Examining the content of documents written in English, it can be seen that the only translation found is "time of flight spectrometer".

Two more steps are needed. They have been tested in a prototype but not at this time included in the commercial product that is used in our applications and for this test.

The first one consist of a feedback to eliminate the translations that are not in the best documents. In our example
“stage, stroke, weather, beat” can be eliminated as translation of "temps"

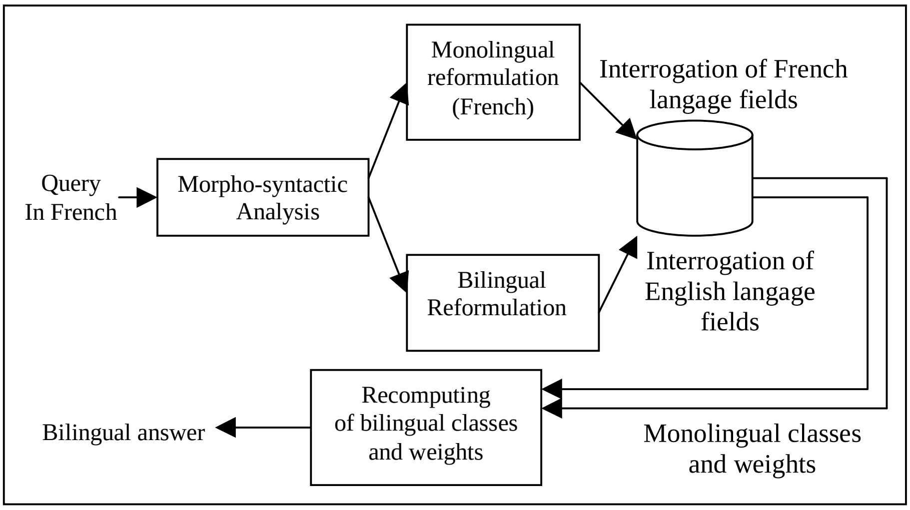
The second one is the use of a second reformulation step using monolingual target lexical knowledge. In our case, this reformulation can infer, for example, "spectrometry" from "spectrometer". This will give a better answer to the query.

4 Accessing Mixed Language Databases

The architecture developed as part of the EMIR project is not sufficient to access mixed language databases. The last version of SPIRIT is now able to process mixed language databases, provided that the information in different languages is kept in different fields. With this version, each language is correctly indexed. But it is necessary to enter separate queries in both languages and manually mix the results.

Because the preceding procedure is “user unfriendly”, we have developed a special crosslingual interrogation system, that can give mixed language results.
A query in one language is used to retrieve data from the part of the database which is in the same language. In parallel, the same query can be used to access the part of the database which is in the other language. This is performed through the crosslingual tools of the system.

Fig. 2. Crosslingual Interrogation



A major point is that the system keeps the link between inferred words and the original query words. Consequently, it is possible to merge the results of the two queries.

Each class of documents is characterized by a particular intersection with the query. In relation to the preceding remark this intersection can be expressed with the original query words even for querying in the other language. It is thus easy to merge the classes of intersections between the two results. In some cases (for a document with more than one language), some classes disappear and some others are created.

For example :
With a query like “gestion des déchets nucléaires” (management of nuclear wastes), the intersection with the English part of the document can be “déchets nucléaires” (nuclear wastes) and with the French part “gestion des déchets” (management of wastes). By merging the results, a new class can appear to place this document, now characterized by “gestion des déchets nucléaires”, and some classes in which this document was can disappear.

If in the merging process the first problem to be solved is the computation of new classes of intersection, an other problem lies in the weighting of words. As weights are calculated for each language, we have to compute a new weight corresponding to the occurrence of the same concept in different languages.

In the case of SPIRIT, the weight of words is linked to the number of documents they appear in.

It is not realistic to aim at a global weighting of the bilingual database since there is no link between words in context in a given language and their correct translation in an other language during the indexing process. Only in the context of a query can the ambiguity be solved.

The solution we have chosen for this problem of weight recomputing is to consider the full answer to the query as a database. In this case it is possible with the same algorithm to merge the classes and recompute a common weight, because we have information on the number of documents a concept occurs in its different linguistic forms.
The experiments done show that this palliative solution gives good results.

Example of use on the CEA's publication database
Let us give an example based upon Intranet querying of the bibliographic database of publications issued by CEA’s scientists.

La conservation des aliments par irradiation.

Result (In the class definition, the comma stands for the AND operator and the hyphen means a dependency relation between two words appearing in both the query and the documents) :

1st class : *conservation-aliments, irradiation***,** 2 documents

- Food preservation by irradiation : a brief introduction
- Le traitement ionisant des aliments..

2nd class : *aliments, irradiation***,** 2 documents

- Devenir de l'ionisation des aliments
- Flux des radionucléides dans les productions agricoles suite à un accident nucléaire

The first document of the first class contains :
In French : conservation des aliments,
In English : food preservation, irradiation, radiations

Remark : "conservation des aliments" has inferred "food preservation" by a word-for- word translation followed by a reordering of the words according to transformational rules.
Irradiation has inferred “irradiation” and “radiation”.

But the dependency relation between "conservation" and "irradiation" has not been taken into account in the French parsing, so the relation between "preservation" and "irradiation" has not been recognized in the English part of the document.

The second document of the first class contains :
In French : conservation des aliments, irradiation, rayonnement
In English : radiation

Remark : the word "rayonnement" has been inferred by the French monolingual reformulation.

The first document of the second class contains :
In French : aliments, alimentaires
In English : food irradiation

Remark : because the relation between “aliments” and “irradiation” has not been recognized in the French query, the compound “food irradiation” has not been recognized as such. Consequently, this document, which should have been put into a class characterized by "irradiation des aliments", is in fact inserted into a class only characterized by the co-occurrence of ”aliments" and "irradiation" containing documents on the contamination of food by radionuclides.

In the last case a relevant document has been put into a class characterized by : aliment.

This is due to the fact that the document contains "aliments ionisés" in French and "irradiated food" in English. French reformulation has not inferred "ioniser" from the word "irradiation". As mentioned above , the commercial system only uses a one- step reformulation at the present time. As a result, even if the relation between "irradiation" and "irradiate" is stored in a monolingual English reformulation rule, it is not possible to activate it.

Querying in English approximately gives the same results. Yet, the first document is characterized by the full compound that has been handled appropriately in English.

5 Problems Limiting the Quality of the Results

There are a large number of problems that limit the quality of crosslingual interrogation. Among them are the following :

- Idiomatic expressions that are noncontiguous are not taken into account in the commercial product (verbs with particles in English, German, or idiomatic verbs in French) examples : switch on (EN), abfahren (GE), mettre en œuvre (FR)
- Some compounds of more than 2 parts are not well analyzed in the commercial product. As a result, compounds are not recognized in the same way in the different languages, so the ranking of relevant documents is not the same according to the query language (see example in §4).
- The lack in the commercial product of a multistep reformulation performing a monolingual target reformulation after the bilingual one. (example to obtain "couronne en or" à partir de "golden crown" it is necessary to infer "doré" from "golden" in the bilingual reformulation step. Then "or" from "doré" by a target language monolingual reformulation.

Yet, the main problems are connected with the coverage and consistency of the linguistic (monolingual and bilingual) dictionaries that are used :

- inconsistency of coding between monolingual dictionaries and bilingual ones
- lack of entries
- lack of translations for an existing entry
- lack of compounds that cannot be translated word to word
- lack of symmetry of bilingual dictionaries

This is why we have focused our efforts on quality control and vocabulary coverage in our scope to improve the linguistic resources of our system.

More details can be found in [5].

6 Different Ways to Enhance the Quality of the Linguistic Resources

To enhance the quality of our linguistic resources we have developed programs that control the integrity of the linguistic resources. This means :

- verifying that a source word of a bilingual dictionary is a normalized word produced by the monolingual dictionary,
- verifying that target words of a bilingual dictionary are normalized words produced by the monolingual dictionary,
- verifying that each compound is an idiomatic expression or a compound recognized through syntactic parsing because the normalization is different ("chemin de fer" is an idiomatic normalized expression and the compound "chemin traverse" is the normalized form of "chemins de traverse".

We have also started a year ago a campaign to increase the vocabulary coverage of the dictionaries in our scope, which ranges over a number of fields - from nuclear technology to astrophysics through medicine, chemistry, and semiconductors.
Two main sources have been exploited :

- the vocabulary of our databases
- the vocabulary from the log files resulting from queries to our Intranet and Internet applications.

The vocabulary has been extracted and compared to the monolingual dictionaries. Normalized vocabularies from the database indexes have been compared to the bilingual dictionaries.

An automatic extraction of compounds has been achieved to study the problem of idiomatic expression coding and suggest expressions to be translated as a whole.
Example of an extraction of compounds in the French and English part of the database of the publication issued by our scientists :

Occurence number	*French*	*Occurence number*	*English*
140	résultat obtenus	351	cross section
138	temps réel	261	experimental result
125	champ magnétique	219	magnetic field
122	étude expérimental	218	heavy ion
122	acier inoxydable	151	experimental datum
114	simulation numérique	144	stainless steel
113	résultat expérimental	141	good agreement
91	microscopie électronique	141	experimental study
83	ion lourd	122	neutron diffraction
80	film mince	119	magnetic structure
80	couche mince	119	magnetic property
77	réacteur nucléaire	118	swift ion
77	produit de fission	102	room temperature
74	section efficace	98	high temperature
74	réacteur à eau	95	fission product
68	étude théorique	88	nuclear plant
68	joint de grains	87	numerical simulation

It can be easily seen that if the repartition of themes is equivalent between documents in French and in English, word that are translated from one another have comparative rank in the list sorted by frequencies.

From the preceding list, we can infer that "cross section" must be translated globally by "section efficace" because "efficace" is not the translation of "cross".

Another source is the INIS (International Nuclear Information System) Thesaurus. This thesaurus has been used to extract synonyms for monolingual reformulation as well as English- French keyword correspondences for bilingual dictionaries.

The following problems have been encountered :

- Words are in upper case characters without accents for French. It was necessary to add accents and rewrite words into a full upper case-lower case set of characters.
- A number of keywords are in the plural form, which is not compatible with the normalized forms of SPIRIT. For example "JEUNES" is used for "jeunes" = teenagers and "JEUNE" is used for "jeûne" = diet.
- Some relations are of no interest for SPIRIT because they are automatically taken into account by linguistic processing.

Example :
The relation between a broader term and a narrower term when it is a relation between a noun phrase and its head (Example : “filter” and “air filter” or “electric filter”) or the relation between synonyms that are different representations of the same word (Example : “ashing (dry)” is equivalent to “dry ashing”).
Concerning compounds, English-French translations may not be true translations but mere explanations of English terms, so this kind of information cannot be used to build up reformulation rules.

Examples :
Carcinogen → produit chimique cancérogène
Chemosterilant → agent chimique stérilisant
College → établissement d'enseignement supérieur
Forestry → valorisation des produits forestiers

All the translations or synonyms produced by the above processing are manually controlled by our terminologist before addition to the existing dictionaries.

The most difficult problem is to discover compounds that require to be translated as a whole. In order to tackle this problem, we have started a research program a year ago to produce an automatic extractor of compound translations from existing translated texts. This system establishes a correspondence between sentences of both texts. It starts from a sentence and its translation, basing itself upon their syntactic analysis and the links established thanks to the existing dictionary, and proposes links between words or compounds that had not yet been related.

At this time the sentence alignment program has been developed and tested as part of the ARCADE international common evaluation program [10]. Because we have privileged precision, and not recall, our results are not the best, but our approach based upon the use of crosslingual interrogation for alignment shows the best standard deviation. Our approach therefore seems to be more robust and giving a minimum difference between the best results and the worst ones.

We are developing word alignment at the present time.

7 Preliminary Evaluation

Comparative evaluation of effectiveness using a common database and set of queries is a very huge work. That is the reason why it is generally a common investment of the world research community like for TREC (organized by NIST USA) or AMARYLLIS.(TREC-like evaluation on French databases done by the French Speaking Countries University Organisation (AUF))

Getting a quantitative evaluation of effectiveness between the actual application that uses a monolingual indexing in French for both English and French part of the information (ML) and the bilingual indexing and retrieval which is proposed in this paper (BL) , should be of a great interest.

Unfortunately, the amount of work to get a significant statistical result was too important. That is the reason why we will only give some qualitative results.

Because the ML application is in exploitation, we have a log file of real queries from users. We have begun to systematically execute these questions and their translation in the other language using the ML application and the BL application.

The first results show different kind of situations :

- the system gives good translations of the query. In this case and provided the fact that there is few bilingual documents the result of the BL system is much better than the ML one.

For example : query : "food preservation by irradiation" gives three relevant documents at the three first positions in BL. ML give only one relevant document.

The same query in French : "conservation des aliments par irradiation" gives three relevant documents at the three first positions. In ML, 2 relevant documents are in the 2 first positions, the third relevant document is at the 8^{th} position.

The difference of effectiveness is strongly influenced by the proportion of relevant documents that have parts in the both languages or not.

- the query uses an expression that cannot be translated word to word and this expression is not in the bilingual reformulation dictionary. The results of ML and CL are similar.

For example, a query on "courant de Foucault" in French gives the same answer in ML and BL (14 documents) because the translation is not known by the system. The result of "eddy current" is better in BL (18 documents) than in ML (12 documents) only because of the lack of normalization in English which is not the language of indexing for ML.

After introduction of the reformulation "courant de Foucault" <-> "eddy current" with BL, 22 relevant documents are retrieved on the top of the list both with the French and the English query

As a conclusion, the increasing of effectiveness varies but is always positive. With the work done on the updating of bilingual dictionaries especially for compounds, the effectiveness will be better and better.

8 Conclusion

Even if the system at this time is still not perfect, the results are much better than those obtained when indexing documents in only one language. Conditions of use are also greatly improved through asking *one* query and getting *one* list of ranked documents, as opposed to querying in both languages successively and performing a manual merging of ranked results.

The interrogation of the bilingual catalogs and the C.E.A.'s scientists publications is subject to quality test. We hope to be able soon to propose an access to this Intranet application to our users.
Of course, a great deal has still to be done to improve this technology. Part of this improvement has been tested in prototypes, but has not yet been integrated into the commercial product. This involves the introduction of noncontiguous idiomatic expressions, multistep reformulation, and a full recognition of dependency relations.

Further improvements mean lasting efforts to enhance quality, consistency, and exhaustiveness not only of the monolingual dictionaries used for query and document parsing but also of the monolingual and bilingual rules required for the reformulation process.

References

1. About reformulation in full text IRS, F. Debili, C. Fluhr, P Radasoa, Conference RIAO 88, MIT Cambridge, mars 1988, Text adapted for the journal "Information processing and management" Vol. 25, N° 6 1989, pp. 647-657.

2. Automatic text processing. The transformation, analysis and retrieval of information by computer. G. Salton (1989) Addison-Wesley publishing company 1989. 530p

3. "Fully automatic cross-language document retrieval using latent semantic indexing." T.K. Landauer, and M.L. Littman In Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research. Waterloo Ontario 1990.

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

5. 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" dans "AAAI 1997 Spring Symposium Series", 24-26 mars 1997, Stanford University, Californie.

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

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

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

9. EMIR at the crosslingual track of TREC-7, F. Bisson, J. Charron, C. Fluhr, D. Schmit, TREC-6 Conference, 9-11 Novembre 1998, Gaithersburg, Maryland

10. Parallel Text Processing, Jean Véronis scientific editor, Kluwer Academic Publishers, Text Speech and Language technology series, to be published in 1999.