

Textual database lexicon used as a filter to resolve semantic ambiguity

Application on Multilingual Information Retrieval

Charles Raynaud and Christian FLUHR
Institut National des Sciences et Techniques Nucléaires
Centre d'Etudes de Saclay - 91191 Gif sur Yvette - FRANCE
e-mail: charles.raynaud@pt.lu

Abstract

The resolution of semantic ambiguity remains one of the most complex subjects of natural language processing (NLP). In that paper we propose an approach to resolve efficiently that question in the field of multilingual information retrieval (MIR) in full text. The paper describes one foundation of the approach: the choice of the sense. The proposed approach is a general one, it is neither language dependent nor application domain dependent.

We develop here a global system for MIR built on three major components : a morpho-syntactic analyzer, a transfer module from a source language to a target language and information retrieval subsystem. A thorough investigation into the system architecture and transfer mechanisms is submitted bellow.

Keywords

Multilingual Information Retrieval (MIR), Information Retrieval (IR), Natural Language Processing (NLP), Machine Translation (MT), End-User Interface.

Introduction

Many companies are handling textual information delivered on multiple languages. The current case is the simultaneous presence text in the mother tongue language of the country and the scientific information in English. The plurilinguism problem is more important in countries where there are more than one official language such as Belgium, Canada, Switzerland. Access to textual databases containing documents in several languages is necessary for the European Community where industrial and research activities are more and more dispatched on the national territories.

These considerations brought us to design the EMIR project (European Multilingual Information Retrieval) which aims at allowing users to express queries in their mother tongue language and getting answers in the language of the document.

We assume that even if a user can read easily a text he may get some difficulty is choosing good term for the query. Using good query term improves the quality of answers.

The mechanisms we develop in the EMIR project bring closer terms expressed by the user to the terms expressed in the text even if the latter is in another language. EMIR's approach uses semantic knowledge based on transfer lexicon.

EMIR is issued from the SPIRIT [FLUHR 90] original IR system, which has for its query facility two major components : a natural language processor (morphological and semantic level) for the text and the query and a statistical processor allowing the calculus of the intersection between query keywords and the textual database documents keywords.

EMIR system provides solutions for the following situations:

- Query in the mother tongue language and the database is in a language that we can read but we do not know sufficiently the language vocabulary to express precisely the query.
- Database is in a language that the user ignores. EMIR then becomes a taking decision tool for translation. It may be interesting to couple the EMIR system to a machine translation system.
- Query in only one language, a database containing documents in several languages even if the user is fluent in all the languages of the database. The user will get an overall view of the answers. EMIR contributes as a tool to improve searcher's productivity.

EMIR project (ESPRIT II 5312) is ended in March 1994 and the complete prototype with full linguistic and reformulation options is delivered in UNIX platform. A derived product of EMIR project will be commercialized soon in the French market as an option of the SPIRIT system. The product will fit a client server architecture and will run in UNIX and Windows platforms. The prototype had been demonstrated in scientific and industrial international manifestation (like RIAO'91, Cebit'92, ESPRIT 92, EWAIC'93, ESPRIT 93, ISTEAD'94).

Linguistic and reformulation options as multiterms recognition and derived multiterms handling will be discussed in an other paper. Those options enhance the overall system performances. The multiterms handling resolves automatically the semantic ambiguity but it is not sufficient to resolve that ambiguity for all the text, so we will focus here on uniterms.

Description of the monolingual system

In order to describe briefly the IR system we will use the following example to illustrate it.

The question is extracted from the CRANFIELD test database question.

The query

"how does a satellite orbit contract under the action of air drag in an atmosphere in which the scale height varies with altitude ?"

The system treats the query to extract a set of keywords and dependancy relations between them.

Extracted keywords

"satellite, orbit, contract, action, air, drag, atmosphere, scale, height, vary, altitude"

Extracted dependancy relations

"satellite-orbit, orbit-contract, contract-action, action-air, air-drag, drag-atmosphere, atmosphere-scale, scale-height, height-vary, vary-altitude"

The IR system perform a research to compare between the query keyword and the database keywords. This gives the following results :

Class N°	Doc. n°	Keywords
1	548	satellite-orbit, air-drag, scale-height, atmosphere, varies, altitude.
2	613	satellite-orbit-contract, air-drag, atmosphere, varies.
3	617	satellite-orbit, air-drag, scale-height, atmosphere, altitude.
4	616, 622	satellite-orbit, scale-height, air, atmosphere, varies.
5	614	satellite-orbit, air-drag, atmosphere, varies.
6	615	satellite-orbit, air-drag, varies.

7	163	satellite-orbit, drag, atmosphere, altitude.
8	621	scale-height, satellite, orbit, air.

At this step the user can choose the suitable class of documents. The user can also decide of the order to display the relevant documents.

Document 617 When consulting the CRANFIELD database we can get an answer like bellow.

Title determination of upper-**atmosphere** air density profile from **satellite** observations.

Author Groves, g.v.

Reference proc. roy. soc. a. 252, 28-34, 1959 .

Text determination of upper-**atmosphere** air density profile from **satellite** observations. The theory previously developed for the changes in the perigee distance and semi-major axis of a **satellite orbit** due to **air drag** is extended to enable the **air-density** profile/i.e.its relative variation with **height**/to be derived from the motion of the **orbit's** perigee . the solution is first obtained in terms of the change in perigee distance and then in terms of the change in the radius of the earth at the sub-perigee point the **scale height** in the 180 and 220 km **altitude** regions.

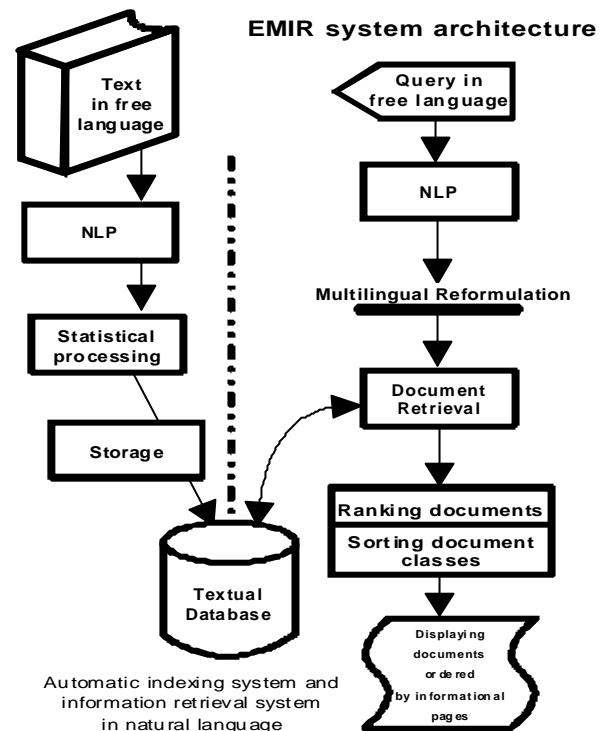
To appreciate more the results we use the 225 queries from set of CRANFIELD queries and we plot the recall/precision curve.

Recall	Precision
10	0,71
20	0,67
30	0,49
40	0,41
50	0,35
60	0,23
70	0,16
80	0,07
90	0,03
Average precision	0,345

For the purpose of multilingual information retrieval we translate the query set in French (225 queries).

System architecture

The system architecture has an automatic indexing component and a consulting component. Both of them are using a NLP system. The NLP system is the same for all languages, just the knowledge base associated to that NLP changes from a language to another.



The NLP query component uses a reformulation system to translate words from the query language into the text language. Reformulation thesauri can help the system to bring closer the query words to the text words by reformulating it with the words used in the database.

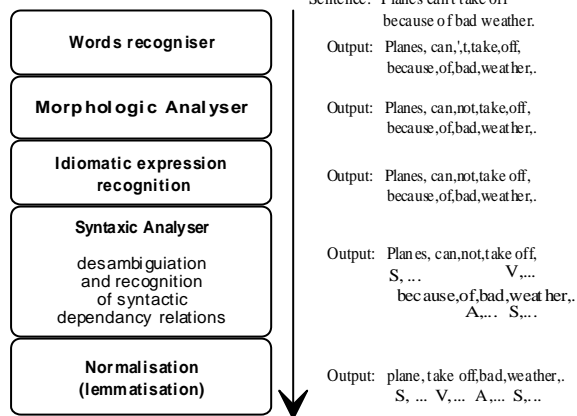
Natural Language Processing

The natural language processing approach makes use of a full form dictionary containing, for every word, all the grammatical characteristics such as part of speech, gender, number, tense. This dictionary is used in morphological analysis. Another dictionary is available for idiomatic

expressions and their grammatical characteristics. The syntactic analysis is based on disambiguation matrices. [Fluhr 77]

The NLP system presents the advantage to be general for all domains, literature, scientific, legal, artistic, etc.

Natural Language Processing
(morpho-syntactic analyser)



The input of the morpho-syntactic analyzer is a query in natural language. Its output is a set of lemmatized significant words, grammatical category for every lemmatized word and dependency relations between keywords.

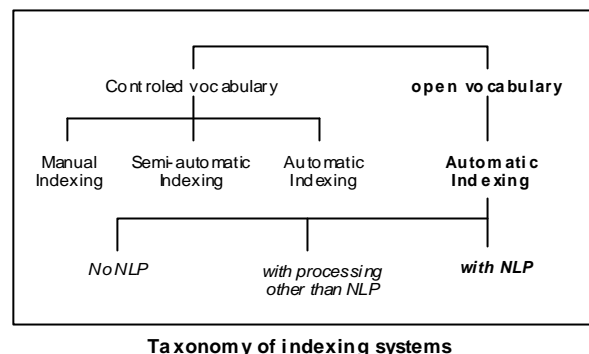
The indexing system will use that information to build inverted lists localizing the different forms of every lemmatised word.

The NLP system allows more than lemmatization of simple words to recognize contiguous and unctiguous idiomatic expressions. This is very important to grantee the choice of the good entry in transfer dictionary. The system is able to recognize a term like "looking forward", the translation of that term is not the same as "look". A NLP without that recognition system will produce "look" instead of "look forward". This kind of expression is frequently used to express the nuance in the language.

NLP integrated in the EMIR system supports the following languages English - French - German - soon Russian, has to be extended to Spanish - Dutch and Arabic.

Which indexing approach is used ?

In order to determine which indexing approach is used for the EMIR system we use the following figure to the taxonomy of indexing.



The bold path shows the indexing method used by the EMIR system. EMIR allows open vocabulary that means that unknown words absent from the language dictionary are allowed in the text. Automatic indexing means that there is no need for operator to intervene during the indexing process. The automatic indexing path includes a natural language processing operation that goes up to normalized level (lemmatisation, morpho-syntactic analyzer). The results of NLP are normalized words (example for English language: the singular of the plural words, and the present tense of verbs).

The inverted list generated from the free text database contains normalized words and points on the location of the with its original form (both derived and not derived words).

Identified idiomatic expressions are integrated in the inverted list in a canonical form.

The study of the inverted list issued from the automatic indexing of many textual databases shows that the number of simple words against the size of database is not linear.

The curve behaves as the equation $y = a \sqrt[b]{x}$

We use for our study the English language Cranfield textual database, 270 judgments of the European Court of Justice and about 1700 claims for patent presented to the European Patent Office in at least three languages we retain for the

experience English and French, we get the following results.

English text	A	B	C
Text length in KBytes	1727	3168	3830
NB of lemmatised words	2924	4328	4831

- A Cranfield textual database
- B 270 judgments of European Court of Justice (ECJ)
- C 1700 claims for patent of European Patent Office (EPO)

$$y \approx 120^{0.444} \sqrt{x} \quad (\text{eq 1})$$

French text	B	C
Text length in KBytes	2000	2500
NB of lemmatised words	5353	6034

$$y \approx 95^{0.509} \sqrt{x} \quad (\text{eq 2})$$

The equations eq 1 and eq 2 were calculated over more than seven different textual databases in each language.

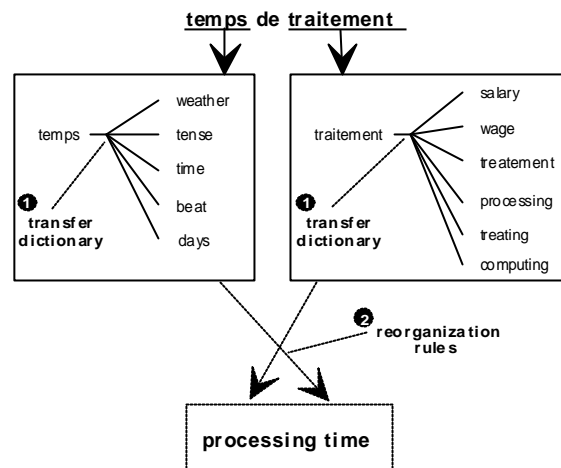
We can observe a stability in the vocabulary for huge texts for both languages. This stability will reduce the effort given for the management of transfert dictionnaires

We can observe also that English and French languages have the same behaviours so our NLP approach is applicable for other languages.

Multilingual reformulation process

The multilingual reformulation process is used when queering the database. The process uses transfer dictionaries and a file containing syntactic reorganization rules to operate. It operates assuming that at the entry there are normalized (lemmatised) words and the inverted list of the textual database is normalized words. [RAYNAUD 94]

The next example shows the principle of the multilingual reformulation.



The Multilingual reformulation motor is running using transfer dictionaries to translate keywords.

At that reformulation stage we use about five reorganization rules so that the order of translated words (two or more) obeys to the syntactic rules of the target language.

At the same stage we recognize what we call multiterms. Multiterms are two or more words that must be translated globally and not by their components, like 'seat belt' or 'air bag'. We treat many cases of multiterms, a uniterm (one word) translated in a multiterm, multiterm translated in a uniterm and multiterm translated in multiterm. We observe that the multiterm detector is a powerful tool to minimize semantic ambiguity and prevent ridiculous translations.

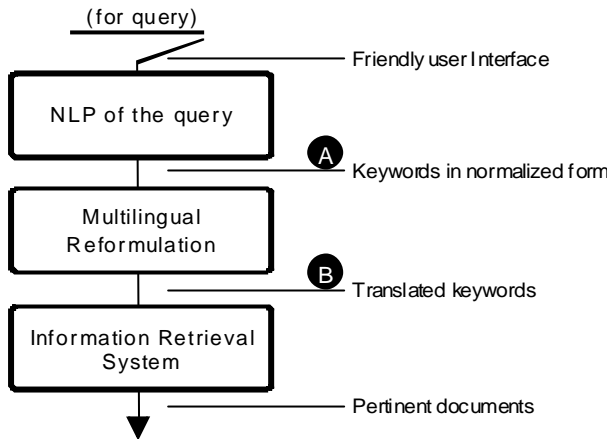
NLP step helps as a consequence to reduce the number of entries in the transfer dictionaries, allows to use part of speech as selection criterion and enhances the possibility of using syntactic dependency relations in the query.

Suggested approach

The suggested approach for the query part of a multilingual IR system is composed of three parts. A natural language processing (NLP) which has in output normalized keywords. A multilingual reformulation motor that generates translations for every keyword. The resulting words are injected in Information Retrieval

System (IR) which will give back the pertinent documents in the database.

Suggested approach



Starting from the two points A and B, we are about to study the effects of filtering possible translations.

MIR system illustration

We use the same query as in the illustration of the monolingual system to illustrate the multilingual one. The CRANFIELD query set had been translated by the French army documentation center.

"Comment une orbite de satellite se contracte-t-elle sous l'action de la traînée de l'air dans une atmosphère dans laquelle l'échelle de hauteur varie avec l'altitude ?"

The NLP step gives the following words :

"orbite, satellite, contracte, action, traînée, air, atmosphère, échelle, hauteur, varie, altitude."

The dependency relations retained are :

"orbite-satellite, satellite-contracte, contracte-action, action-traînée, traînée-air, air-atmosphère, atmosphère-échelle, échelle-hauteur, hauteur-varie, varie-altitude."

The ranked documents resulting from IR are :

Class N°	Doc N°	Keywords
1	617	orbite-satellite, traînée-air-atmosphère, contracte, échelle, hauteur, varie.
2	616	orbite-satellite, air-atmosphère, contracte, échelle, hauteur, varie.
3	548	orbite-satellite, traînée-air, contracte, échelle, hauteur, varie.
4	162	orbite-satellite, hauteur-varie, air.
5	622	orbite-satellite, contracte, action, air, échelle, hauteur, varie.
6	613, 614, 615	orbite-satellite, traînée-air, contracte, varie.
7	618	orbite-satellite, traînée-air, hauteur.

The most relevant document appears with english keywords highlighted as following.

Document 617

Title determination of upper-**atmosphere air** density profile from **satellite** observations.

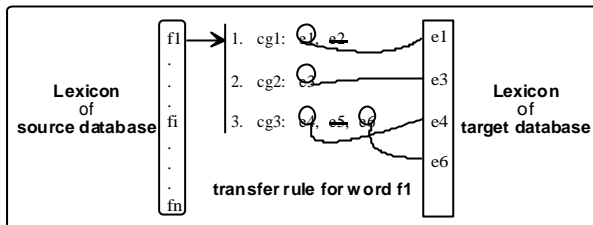
Author Groves, g.v.

Reference proc. roy. soc. a. 252, 28-34, 1959 .

Text determination of upper-**atmosphere air** density profile from **satellite** observations. The theory previously developed for the **changes** in the perigee distance and semi-major axis of a **satellite orbit** due to **air drag** is extended to enable the **air**-density profile/i.e.its relative variation with **height**/to be derived from the motion of the **orbit's** perigee . the solution is first **obtained** in terms of the **change** in perigee distance and then in terms of the **change** in the radius of the earth at the sub-perigee point the **scale height** in the 180 and 220 km **altitude** regions.

Lexicon of textual database used as a natural semantic filter

We prepare our experience using the textual databases ECJ (270 European Court of Justice judgments) and EPO (1700 claims for patent, text offered by the European Patent Office). The texts which are in English and French are the perfect translation of each other.



General principle for the filtering experience

The four databases ECJ English, ECJ French, EPO English and EPO French are automatically indexed. We extract the single words from the inverted list. Extracted words are translated using a transfer dictionary (see description by the end of the paper). For every word we consult the inverted list of the corresponding database in the other language and we verify the presence of translation according to the following system.

In order to determine the effectiveness of lexicon of the textual data base as a filter, we limit the experience to uniterms, multiterms are usually not polysemic.

Filtering ratio

In order to calculate the filtering ratio we will find for each lemmatised word in the source database its possible translations in the target language database using the transfer dictionary. Once we get all the needed transfer rules we can look for each translation in the target database establishing a ratio between the number of available translations in the transfer dictionary and those present in the target database.

Assume that a transfer rule '*rt*' has at its right side *k* translations, $x = ratio(rt)$.

$$ratio(rt) = \frac{\text{nb of translations in target lexicon}}{k}$$

$$f(k, x) \rightarrow \frac{\text{nb of rules having } k \text{ translations and at } x\% \text{ ratio}}{\text{nb of rules having } k \text{ translations}}$$

$$x \in \{10\%, 30\%, 50\%, 70\%, 90\%\}$$

$$f(k, x) = \frac{\sum_{i_k < n_k} d(r_{k,i_k}, x)}{n_k}$$

$$d(r, x) = \begin{cases} 1 & x - 0.1 < ratio(r) < x + 0.1 \\ 0 & \text{other} \end{cases}$$

Average is calculated using the formula

$$average = \frac{\sum_{k \in K} f(k, x)}{\text{nb of elements in } K}$$

K is the set of entries in the following table.

In the table bellow we ranked the transfer rules by number of translations in it and separate by an interval of ratio (of 20%).

The table is the result of the experience done on the 270 judgments of the European Court of Justice lexicons. The French lexicon was the source one and the English lexicon was the target one. To interpret this table, if we take the 5th line, this means that we have 372 rules getting 5 translations; 75 rules got one translation over five, 107 rules got two translations over five, etc.

NB of translations.	NB of rules	at 20%	at 40%	at 60%	at 80%	at 100%
1	956	----	----	----	----	100%
2	865	----	61,8%	----	----	38,2%
3	695	39,4%	----	38,6%	----	22,0%
4	486	34,0%	29,8%	25,9%	----	10,3%
5	372	20,2%	28,8%	28,0%	16,9%	6,2%
6	252	40,5%	23,8%	21,8%	11,1%	2,8%
7	187	32,6%	42,8%	13,4%	8,6%	2,7%
8	128	49,2%	21,1%	21,1%	5,5%	3,1%
9	103	40,8%	35,0%	16,5%	6,8%	1,0%
10	69	36,2%	33,3%	23,2%	7,2%	----
11	44	50,0%	31,8%	11,4%	6,8%	----
12	28	35,7%	46,4%	17,9%	----	----
13	18	55,6%	27,8%	16,7%	----	----
14	11	36,4%	54,5%	9,1%	----	----
15	16	56,3%	25,0%	18,8%	----	----

16	6	33,3%	50,0%	16,7%	---	---
17	6	50,0%	16,7%		---	---
18	7	57,1%	28,6%	14,3%	---	---
19	6	83,3%	16,7%		---	---
21	1	100,0%			---	---
23	1	100,0%			---	---
24	1	100,0%			---	---
Average		43,7%	31,4%	17,5%	3,9%	3,3%
Cumulated average		43,7%	75,1%	92,6%	96,7%	100%

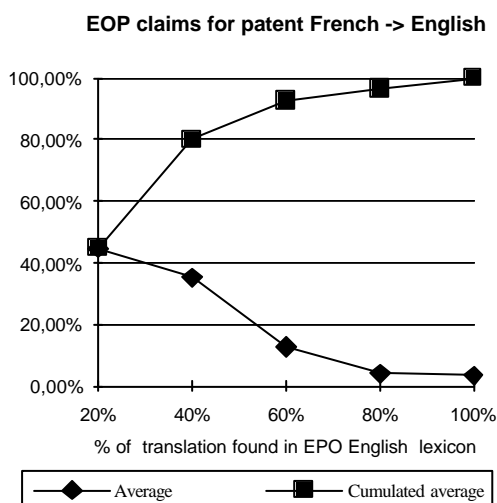
The following graph shows the results of the same experience as before applied to the EOP claims of patents and the ECJ judgements.

The major part of the translations is rejected so in more than 40% of the cases we got one over five translations and in 80% of the cases we got two over five translations.

These results had been verified for the two databases in both ways English to French and French to English. The graphs below illustrate the results.

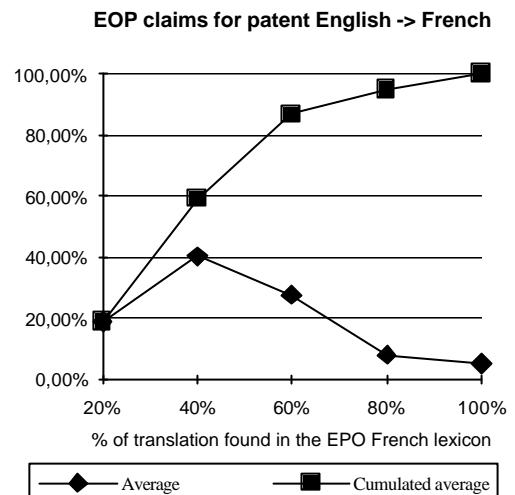
1 French lexicon of EPO claims for patent database as source and English lexicon as target.

	20%	40%	60%	80%	100%
Average	44,8%	35,2%	12,6%	4,0%	3,4%
Cumulated average	44,8%	80,0%	92,6%	96,6%	100,0%



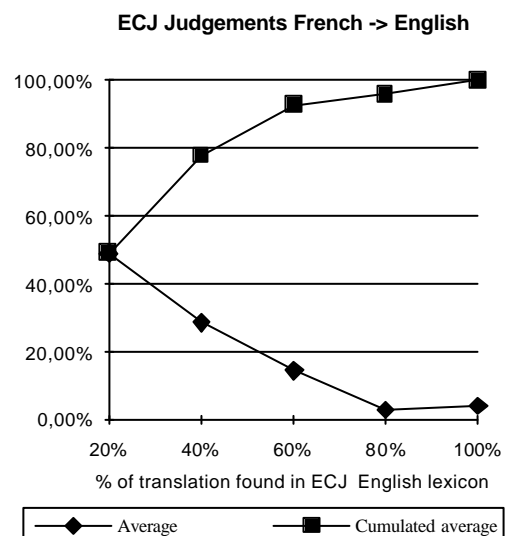
2 English lexicon of EPO claims for patent database as source and French lexicon as target.

	20%	40%	60%	80%	100%
Average	18,9%	40,3%	27,7%	8,0%	5,1%
Cumulated average	18,9%	59,2%	86,9%	94,9%	100,0%



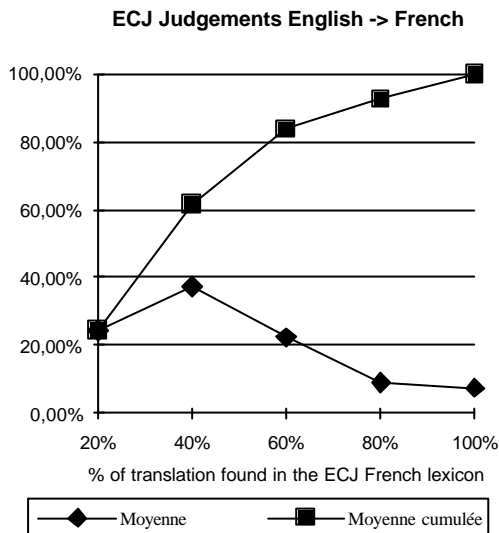
3 French lexicon of ECJ judgements database as source and English lexicon as target.

	20%	40%	60%	80%	100%
Average	49,2%	28,9%	14,8%	3,0%	4,1%
Cumulated average	49,2%	78,1%	92,9%	95,9%	100,0%



4 English lexicon of ECJ judgments database as source and French lexicon as target.

	20%	40%	60%	80%	100%
Average	24,3%	37,2%	22,4%	8,9%	7,4%
Cumulated average	24,3%	61,4%	83,8%	92,7%	100,0%



The database of judgments treats multiple subjects in several domains. Judgments on financial, industrial, import-export, agriculture, foodstuff affairs are treated in that database. The claims for patent database are treating mainly of electronics. The behavior of the graphs concerning the same language is similar even when changing the database.

Conclusion

The lexicon of the textual database appears as a first step filter to resolve semantic ambiguity. This approach goes through the problem of deciding which translation to use. Another part of the EMIR system accurate the localization of documents containing the exact translations for the query terms.

The approach allows to integrate specialized dictionaries so we can enlarge the semantic field of a word without generating noise in the IR system. The noise can come by adding, to the transfer rule at the right side, words which are very polysemic like 'do', 'have'. So a special

attention has to be brought to the transfer rule production.

NLP process helps the IR system giving it a friendly interface, helps to reduce the number of entries in the transfer dictionaries to the lemmatised forms and allow the usage syntactic reorganization rules.

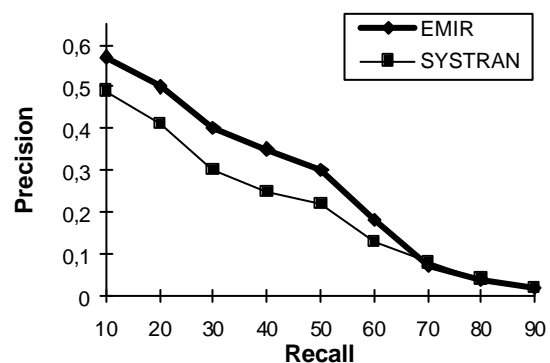
The retrieval performances [Salton 83] had been studied using the CRANFIELD test set for two approaches. The first one is what we suggest, the second one is to translate queries to the data base language using SYSTRAN translation system. The performances were as below.

The curve compares between two approaches. The first one the EMIR system approach, the second one is to translate the query by SYSTRAN system and query using the translation.[RAYNAUD 94]

The performance of the EMIR system can be enhanced by integrating the contiguous and unctiguous expression recognizer; a lot of linguistic data has to be prepared before that integration.

The most we enhance the quality of results of NLP the best we get IR performance. Many functionalities can be added to NLP as multiple dependency relations can enhance the precision of IR system.

Comparison between bilingual query using EMIR and monolingual query after translation by SYSTRAN



The approach had been verified to query in German and to query a German language database. EMIR system allows now to treat completely (indexing and consulting) the three languages English, French and German.

Bibliography

[Croft 91]

W.B. Croft, H.R. Turtle, D.D. Lewis. The use of phrases and structured queries in information retrieval. SIGIR'91, pp. 2-41, 1991.

[Fluhr, RAYNAUD 90]

C. Fluhr and C. RAYNAUD. "Accès multilingue aux bases de données en texte intégral" - Leningrad, URSS, mai 1990. [IAEA-SM-317/16]

[Fluhr, RAYNAUD 90]

C. Fluhr, C. RAYNAUD. "Full text databases as lexical semantic knowledge for multilingual interrogation and machine translation" CEI Moscow, East West Artificial Intelligence conference, EWAIC'93, 7-9 September 1993.

[Fluhr 77]

C. Fluhr. "Algorithmes à apprentissage et traitement automatique des langues." Thèse d'état N° 1863. Université de Paris-Sud. Orsay, 15 juin 1977.

[Fluhr 90]

C. Fluhr "Multilingual access to full text databases" — International A.I. symposium 90 Nagoya, November 14-16, 1990, Nagoya, Japan, pp. 107-110

[Fox 88]

E.A. Fox, G.L. Nunn, W.C. Lee. "Coefficients for combining concept classes in a collection." SIGIR'88, pp. 291-307, 1988.

[Fox 90]

"Virginia Disc One" Nimbus Records, Inc. Published by Virginia Polytechnic Institute and State University Press, Editor, Project Director, Principal Investigator Edward A. Fox, Dept. of Computer Science. CDROM. 1990.

[Harman 92]

D. Harman. "Evaluation issues in information retrieval." Information Processing & Management Vol. 28, No. 4, pp. 439-440, 1992.

[Kristensen 93]

J. Kristensen. "Expanding end-users' query statements for free text searching with a search-aid thesaurus." Information Processing & Management Vol. 29, No. 6, 1993, pp. 733-744.

[Molto 93]

M. Molto. "Improving full text search performance through textual analysis" Information Processing & Management Vol. 29, No. 5, pp. 615-632, 1993.

[RAYNAUD, al. 91]

C. RAYNAUD, F. Foussier, C. Fluhr. "Multilingual access for textual data base." RIAO'91 Barcelona 3-5 April 1991. pp. 475-489.

[RAYNAUD, Fluhr 94]

C. RAYNAUD, C. Fluhr. "An intelligent system for multilingual information retrieval" Egypt, Cairo, IASTEAD'94, 26-29 December 1994.

[RAYNAUD 94]

C. RAYNAUD. "Vers l'accès multilingue en langage naturel aux bases de données textuelles." Ph.D. : 3081, Février 1994 de l'Université Paris-Sud d'Orsay, IEF - Institut d' Electronique Fondamentale. 245 p.

[Salton, Gill 83]

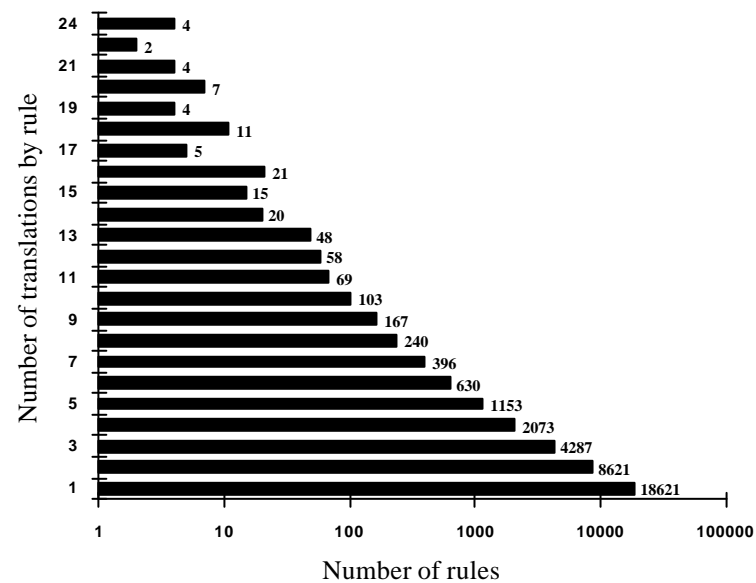
G. Salton, M.J. McGill. "Introduction to modern information retrieval." McGraw-Hill, pp. 157-198, 1983

**English → French transfer dictionary
(Number of entries = 30633)**

Number of translations by rule	Number of rules containing the corresponding number of translations
1	18621
2	8621
3	4287
4	2073
5	1153
6	630
7	396
8	240
9	167
10	103
11	69
12	58
13	48
14	20
15	15
16	21
17	5
18	11
19	4
20	7
21	4
23	2
24	4
Total number of transfer rules in the transfer dictionary	36559

Average	2,117973	100%
from 2 to 24	3,278514	49%
from 3 to 24	4,461521	25%
from 4 to 24	5,707157	14%
from 5 to 24	6,903956	8%
from 6 to 24	8,120842	5%
from 7 to 24	9,258943	3%

**Frequency distribution of transfer rules depending on number of translations
English to French dictionary**



**French → English transfer dictionary
(Number of entries = 30300)**

Number of translations by rule	Number of rules containing the corresponding number of translations
1	14643
2	8109
3	4469
4	2431
5	1384
6	810
7	479
8	266
9	190
10	119
11	87
12	50
13	30
14	22
15	26
16	8
17	6
18	9
19	8
21	1
23	1
24	1
Total number of transfer rules in the transfer dictionary	33153

Average	2,337827	100%
from 2 to 24	3,396164	56%
from 3 to 24	4,484664	31%
from 4 to 24	5,603169	18%
from 5 to 24	6,716366	11%
from 6 to 24	7,838450	6%
from 7 to 24	8,977811	4%

**Frequency distribution of transfer rules depending on number of translations
French to English dictionary**

