A Deep Linguistic Analysis for Cross-language Information Retrieval

Nasredine Semmar, Meriama Laib, Christian Fluhr

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue
Laboratoire d'Intégration des Systèmes et des Technologies
Commissariat à l'Energie Atomique
Centre de Fontenay aux Roses
18, rue du Panorama
92265 Fontenay-aux-Roses, France
nasredine.semmar@cea.fr, meriama.laib@cea.fr, christian.fluhr@cea.fr

Abstract

Cross-language information retrieval consists in providing a query in one language and searching documents in one or different languages. These documents are ordered by the probability of being relevant to the user's request. The highest ranked document is considered to be the most likely relevant document. The LIC2M cross-language information retrieval system is a weighted Boolean search engine based on a deep linguistic analysis of the query and the documents. This system is composed of a linguistic analyzer, a statistic analyzer, a reformulator, a comparator and a search engine. The linguistic analysis processes both documents to be indexed and queries to extract concepts representing their content. This analysis includes a morphological analysis, a part-of-speech tagging and a syntactic analysis. In this paper, we present the deep linguistic analysis used in the LIC2M cross-lingual search engine and we will particularly focus on the impact of the syntactic analysis on the retrieval effectiveness.

1. Introduction

The role of an information retrieval system is to provide relevant documents according to the submitted query, and to locate as precisely as possible the most informational parts of these documents. In order to achieve this goal, conventional information retrieval tools match the query representation against a single document representation to produce a ranked document list (Kammoun et al., 2005). For cross-lingual information retrieval, the goal is to find relevant documents that are in a different language from that of the query (Grefenstette, 1998).

In this paper, we present the impact of a deep linguistic analysis of documents and queries on the retrieval effectiveness of a cross-lingual information retrieval system.

We present in section 2 the main components of the LIC2M cross-lingual search engine, in particular, we will focus on the syntactic analyzer of the linguistic processing. In section 3, a prototype of the LIC2M cross-lingual search engine developed during the European project ALMA (Arabic Language Multilingual Application) is described. We discuss in section 4 results obtained after submitting two runs of questions (one with activating the syntactic analyzer during the linguistic analysis and the other without activating the syntactic analyzer). Section 5 concludes our study and presents our future work.

2. LIC2M Cross-lingual Search Engine

The LIC2M cross-lingual information retrieval system (Besançon et al., 2003) can manage textual database in Arabic, Chinese, English, French, German and Spanish and is composed of the following modules:

- A linguistic analyzer which includes a morphological analyzer, a part-of-speech tagger and a syntactic analyzer. The linguistic analyzer processes both documents to be indexed and queries to produce a set of normalized lemmas, a set of named entities and a set of nominal compounds with their morpho-syntactic tags.
- A statistical analyzer, that computes for documents to be indexed concept weights based on concept database frequencies.
- A comparator, which computes intersections between queries and documents and provides a relevance weight for each intersection.
- A reformulator, to expand queries during the search. The expansion is used to infer from the original query words other words expressing the same concepts. The expansion can be in the same language (synonyms, etc.) or in different language.
- An indexer to build the inverted files of the documents on the basis of their linguistic analysis and to store indexed documents in a database.
- A search engine which retrieves the ranked, relevant documents from the indexes according to the corresponding reformulated query and then merges the results obtained for each language taking into account the original words of the query and their weights in order to score the documents.

2.1. Linguistic Analysis

The linguistic analysis of the LIC2M cross-language information retrieval system is built using a traditional architecture (Figure 1) and is composed of several modules which are common for all the languages with some variation for specific languages. This analyzer proceeds as follows:

The Tokenizer separates the input stream into a graph of words. This separation is achieved by an automaton developed for each language and a set of segmentation rules.

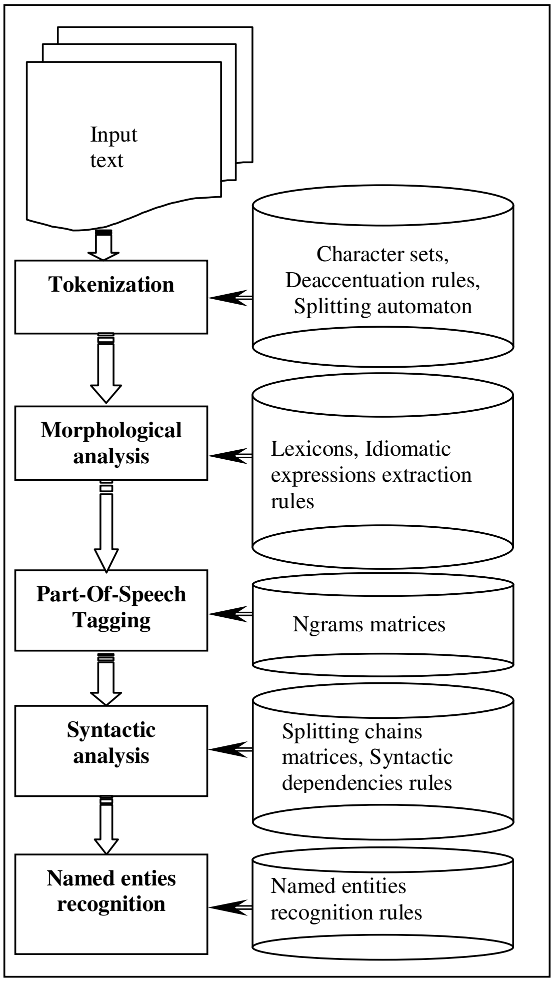


Figure 1: Architecture of the LIC2M Linguistic Analyzer

The Morphological analyzer searches each word in a general dictionary. If this word is found, it will be associated with its lemma and all its morpho-syntactic tags. If the word is not found in the general dictionary, it is given a default set of morpho-syntactic tags based on its typography. For Arabic, we added to the morphological analyzer a new processing step: a Clitic stemmer (Larkey et al., 2002) which splits agglutinated words into proclitics, simple forms and enclitics.

At this point in the processing, an idiom will not be found in the dictionary since we had decided not to include in the lexicon sequence of known words that act as a single unit.

The role of Idiomatic Expressions recognizer is to detect idiomatic expressions and to consider them as single words for the rest of the processing. Idiomatic expressions are phrases or compound nouns that are listed in a specific dictionary. The detection of idiomatic expressions is performed by applying a set of rules that are triggered on specific words and tested on left and right contexts of the trigger.

These rules can recognize contiguous expressions as the "white house" in English, la "maison blanche" in French or "الـــبَـــيْـــت الأبْـــيَض" in Arabic. Non-contiguous expressions such as phrasal verbs in English: "switch…on" or "tomber vaguement dans les pommes" in French are recognized too.

The next step in our linguistic analysis, after idiomatic expressions recognition, is part-of-speech tagging which consists in assigning to a word its disambiguated morpho-syntactic tag in the sentential context in which the word is used.

Our Part-Of-Speech (POS) tagger searches valid paths through all the possible tags paths using attested trigrams and bigrams sequences. The trigram and bigram matrices are generated from a manually annotated training corpus. They are extracted from a hand-tagged corpora of 13 200 words for Arabic, 239 000 words for English and 25 000 words for French. If no continuous trigram full path is found, the POS tagger tries to use bigrams at the points where the trigrams were not found in the matrix. If no bigrams allow to complete the path, the word is left undisambiguated. The accuracy of the LIC2M part-of-speech tagger is around 91% for Arabic, 93% for English and 94% for French (Semmar et al., 2005).

After part-of-speech tagging, a Syntactic analyzer is used to split word graph into nominal and verbal chain and recognize dependency relations (especially those within compounds) by using a set of syntactic rules. We developed a set of dependency relations to link nouns to other nouns, a noun with a proper noun, a proper noun with the post nominal adjective and a noun with a post nominal adjective. These relations are restricted to the same nominal chain and are used to compute compound words.

For example, in the nominal chain "use of nuclear energy", the syntactic analyzer links words as following (Figure 2) because "use" and "energy" are tagged as nouns and "nuclear" as an adjective.

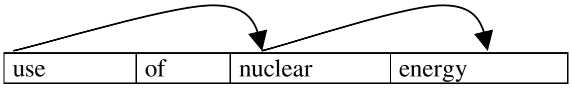


Figure 2: Dependency relations recognition

These relations are used to compute three compound words that are normalized as "use_nuclear_energy", "nuclear_energy", and "use_energy".

The next step in the LIC2M linguistic analyzer, after syntactic analysis, is Named Entity recognition (Abuleil & Evans, 2004) using name triggers (e.g., President, lake, corporation, etc.). Specific named entities extraction is done by using a same method as that used to recognize idiomatic expressions.

For example, in the sentence " British forces, the second largest coalition force in Iraq with just over 8,000 troops, are also expected to start withdrawing in 2006.". "British forces" is recognized as a named entity of type "Organization", "Iraq" is recognized as a "Location", "8000" is recognized as a measure and "2006" is recognized as "Time".

2.2. Statistical Analysis

The role of the statistical analysis is to attribute to each word or a compound word a weight according to the information it provides to choose the document relevant to the query. The weight is maximum for words appearing in one single document and minimum for words appearing in all the documents. This weight is used by the comparator to compare intersection between query and documents containing different words.

The LIC2M search engine uses a weighted Boolean model, in which documents are grouped into classes characterized by the same set of concepts. The classes constitute a discrete partition of the database. For example, if the query is "nuclear waste" on a database containing only texts on nuclear plants, the statistical model indicates that documents containing the compound word "nuclear waste" are more relevant than documents containing the words "nuclear" and "waste". Documents containing the words "nuclear" and "waste" are more relevant than documents containing the word "waste" and documents containing the word "waste" are more relevant than documents containing the word "nuclear".

2.3. Query Reformulation

Reformulation consists in inferring new words from the original query words according to a lexical semantic knowledge database (synonyms, etc.). The reformulation can be used to increase the quality of the retrieval in a monolingual interrogation (Debili et al., 1989). It can also be used to infer words in other languages. The query terms are translated using bilingual dictionaries. Each term of the query is translated into several terms in target language. The translated words form the search terms of the reformulated query. The links between the search terms and the query concepts can also be weighted by a confidence value indicating the relevance of the translation.

2.4. Indexing and Search

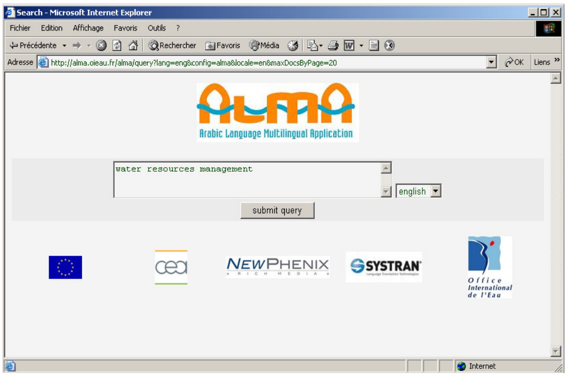
The LIC2M Search Engine Indexer builds the inverted files of the documents on the basis of their linguistic analysis: one index is built for each language of the document collection. This Indexer builds separate indexes for each language.

The LIC2M Search Engine retrieves the ranked, relevant documents from the indexes according to the corresponding reformulated query and merges the results obtained for each language taking into account the original terms of the query (before reformulation) and their weights in order to score the documents.

3. LIC2M Search Engine User Interface

The LIC2M linguistic analyzer produces a set of normalized lemmas, a set of named entities and a set of nominal compounds. These results are used as a front-end for the LIC2M cross-lingual search engine which is visible online at a third party site: http://alma.oieau.fr (Semmar & Fluhr, 2004).

The user can enter a query in natural language and specify the language to be used. In the example of Figure 3, the user entered the query "water resources management" and selected English as the language of the query.


Search - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse
OK
Liens
ALMA
Arabic Language Multilingual Application
submit query
NEWPHENIX
RICH MEDIA
SYSTRAN
cea
Office International de l'Eau
Internet

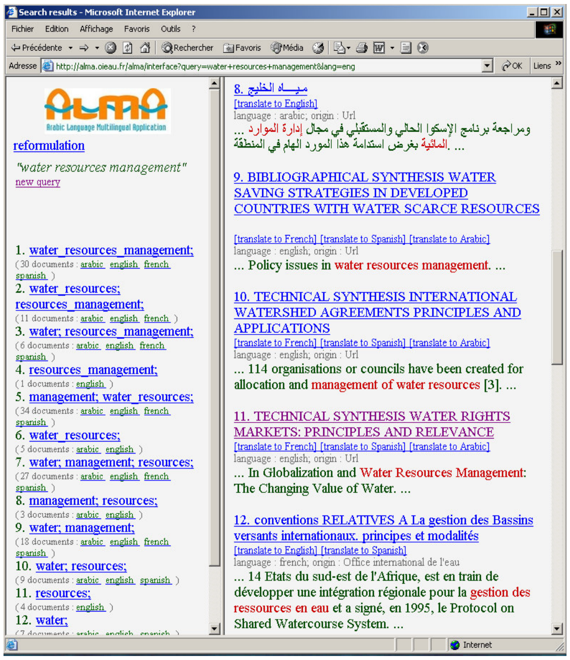
http://alma.oieau.fr/alma/query?lang=eng&config=alma&locale=en&maxDocsByPage=20

water resources management

english

Figure 3: User interface for querying the database

Relevant documents are grouped into classes characterized by the same set of concepts as the query contains (Figure 4). The query term "water_resources_management" is a term composed of three words: water, resources and management. This compound word, its derived variants and their sub elements are reformulated into Arabic and French, and submitted to indexed versions of documents in each of these languages (as well as against English documents).


Search results - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse
OK
Liens
ALMA
Arabic Language Multilingual Application
Internet

http://alma.oieau.fr/alma/interface?query=water+resources+management&lang=eng

reformulation

"water resources management"
new query

1. water_resources_management;
(30 documents : arabic english french spanish)
2. water_resources; resources_management;
(11 documents : arabic english french)
3. water; resources_management;
(6 documents : arabic english french spanish)
4. resources_management;
(1 documents : english)
5. management; water_resources;
(34 documents : arabic english french spanish)
6. water_resources;
(5 documents : arabic english)
7. water; management; resources;
(27 documents : arabic english french spanish)
8. management; resources;
(3 documents : arabic english)
9. water; management;
(18 documents : arabic english french spanish)
10. water; resources;
(9 documents : arabic english spanish)
11. resources;
(4 documents : english)
12. water;
(7 documents : arabic english spanish)

8. مـيــــاه الخليج
[translate to English]
language : arabic; origin : Url
... ومراجعة برنامج الإسكوا الحالي والمستقبلي في مجال إدارة الموارد
المائية بغرض استدامة هذا المورد الهام في المنطقة. ...

9. BIBLIOGRAPHICAL SYNTHESIS WATER SAVING STRATEGIES IN DEVELOPED COUNTRIES WITH WATER SCARCE RESOURCES
[translate to French] [translate to Spanish] [translate to Arabic]
language : english; origin : Url
... Policy issues in water resources management. ...

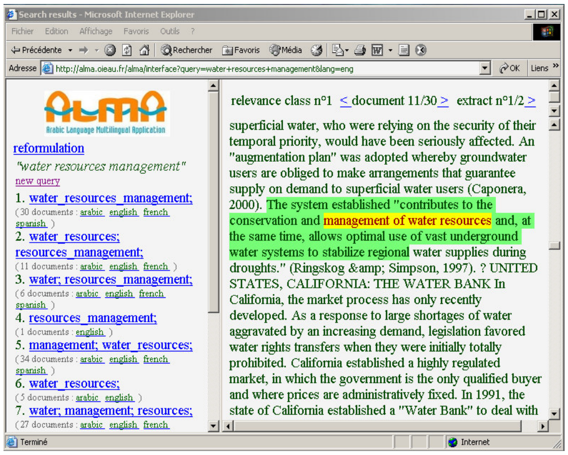
10. TECHNICAL SYNTHESIS INTERNATIONAL WATERSHED AGREEMENTS PRINCIPLES AND APPLICATIONS
[translate to French] [translate to Spanish] [translate to Arabic]
language : english; origin : Url
... 114 organisations or councils have been created for allocation and management of water resources [3]. ...

11. TECHNICAL SYNTHESIS WATER RIGHTS MARKETS: PRINCIPLES AND RELEVANCE
[translate to French] [translate to Spanish] [translate to Arabic]
language : english; origin : Url
... In Globalization and Water Resources Management: The Changing Value of Water. ...

12. conventions RELATIVES A La gestion des Bassins versants internationaux. principes et modalités
[translate to English] [translate to Spanish]
language : french; origin : Office international de l'eau
... 14 Etats du sud-est de l'Afrique, est en train de développer une intégration régionale pour la gestion des ressources en eau et a signé, en 1995, le Protocol on Shared Watercourse System. ...

Figure 4: Search results user interface

Terms of the query (or the expansion of these terms) which are found in the retrieved documents are highlighted as illustrated in Figure 5.


Search results - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://alma.oieau.fr/alma/interface?query=water+resources+management&lang=eng
OK
Liens
Arabic Language Multilingual Application
Terminé
Internet

reformulation

"water resources management"

new query

1. water_resources_management;
(30 documents : arabic english french spanish)
2. water_resources; resources_management;
(11 documents : arabic english french)
3. water; resources_management;
(6 documents : arabic english french spanish)
4. resources_management;
(1 documents : english)
5. management; water_resources;
(34 documents : arabic english french spanish)
6. water_resources;
(5 documents : arabic english)
7. water; management; resources;
(27 documents : arabic english french

relevance class n°1 < document 11/30 > extract n°1/2 >

superficial water, who were relying on the security of their temporal priority, would have been seriously affected. An "augmentation plan" was adopted whereby groundwater users are obliged to make arrangements that guarantee supply on demand to superficial water users (Caponera, 2000). The system established "contributes to the conservation and management of water resources and, at the same time, allows optimal use of vast underground water systems to stabilize regional water supplies during droughts." (Ringskog &amp; Simpson, 1997). ? UNITED STATES, CALIFORNIA: THE WATER BANK In California, the market process has only recently developed. As a response to large shortages of water aggravated by an increasing demand, legislation favored water rights transfers when they were initially totally prohibited. California established a highly regulated market, in which the government is the only qualified buyer and where prices are administratively fixed. In 1991, the state of California established a "Water Bank" to deal with

Figure 5: Highlighting query terms in retrieved documents

4. Experimental Results and Discussion

The LIC2M search engine has been tested on a multilingual corpora provided by partners of ALMA project. This base contains for each language (Arabic, English and French) 50 non-parallel documents related to sustainable development, water and eco-tourism. The test consisted to submit two runs of questions in Arabic, English and French (one with activating the syntactic analyzer during the linguistic analysis and the other without activating the syntactic analyzer) and to compare the retrieved documents.

Table 1 illustrates some classes corresponding to the query “water resources management”. Query terms of classes 1 to 6 correspond to compound words computed by the syntactic analyzer and query terms of class 7 are simple words.

Class number	Query terms	Number of retrieved documents
1	water_resources_management	8 (Arabic) 3 (English) 17 (French)
2	resources_management, water_resources	5 (Arabic) 1 (English) 5 (French)
3	water; resources_management	1 (Arabic) 1 (English) 2 (French)
4	resources_management	1 (English)
5	Management, water_resources	19 (Arabic) 5 (English) 9 (French)
6	water_resources	4 (Arabic) 1 (English)
7	Water, management, resources	4 (Arabic) 14 (English) 6 (French)

Table 1: First classes corresponding to the query "water resources management".

As expected, syntactic analysis improves both monolingual and bilingual reformulation. For example, three new compound words are inferred: "إدارة الـمـوارد الـمـائـيـة" in Arabic, "management of water resources" in English, and "gestion équilibrée de la ressource en eau" in French. Moreover, the French compound word contains the word "équilibrée" which is a new word since the original English query does not contain a translation of this word.

In another query, the English compound word "drinking water" is translated into the Arabic compound word "الـمـيـاه الصـالحـة للشـرب" which is composed of three words instead of two.

5. Conclusion and Future Work

The results we obtained show that it is possible to improve cross-lingual retrieval effectiveness by using a robust and deep linguistic analysis. This robustness is made possible thanks to the syntactic analysis. On the other hand, using this deep analysis allowed the search engine to have a user interface which presents the relevant documents at the beginning of the list of retrieved documents. To confirm these results, we are currently working on a large test database and in the same time we are improving the syntactic analysis.

6. Acknowledgments

This research work was supported by the EU EURO-MED programme project ALMA (Arabic Language Multilingual Application).

7. References

Kammoun, H., Lamirel, J.C., Ben Ahmed, M. (2005). Machine-Learning applied to Queries and Documents for an Adaptive and Evolutive Information Retrieval. In *Proceedings of the International Conference on Machine Intelligence,* pp. 537-545.

Grefenstette, G. (1998). Problems and approaches to Cross Language Information Retrieval. In *Proceedings of the ASIS Annual Meeting,* pp. 35-143.

Besançon, R., De Chalendar, G., Ferret, O., Fluhr, C., Mesnard, O., Naets, N. (2003). The LIC2M’s CLEF 2003 system. In *Working Notes for the CLEF 2003 Workshop*, pp. 83-92.

Larkey, L. S., Ballesteros, L., Connell, M. E. (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 275-282.

Semmar, N., Elkateb-Gara, F., Laib, M., Fluhr, C. (2005). A Cross-language information retrieval system based on linguistic and statistical approaches. In *Proceedings of the Deuxième Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la Langue,* pp. 114-125.

Abuleil, S., Evens, M. (2004). Named Entity Recognition and Classification for Text in Arabic. In *Proceedings of IASSE,* pp. 89-94.

Debili, F., Fluhr, C., Radasoa, P. (1989). About reformulation in full text IRS. In *Informtation processing and management,* pp. 647-657.

Semmar, N., Fluhr, C. (2004). Multilingual Search Engine implementation. In *Final report of ALMA project, EURO-MED programme DG XIII, Commission of the European Union.*