A Cross-language information retrieval system based on linguistic and statistical approaches[*]

Nasredine Semmar, Faïza Elkateb-Gara, Meriama Laib and Christian Fluhr
Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue
DRT/LIST/DTSI/SCRI/LIC2M
Commissariat à l'Energie Atomique
Centre de Fontenay aux Roses
18, rue du Panorama
BP 6
92265 Fontenay aux Roses, France
Tél.: +33 1 46 54 80 15
Fax: +33 1 46 54 75 80
nasredine.semmar@cea.fr, faiza.gara@cea.fr, meriama.laib@cea.fr, christian.fluhr@cea.fr

Abstract

As the number of non-English documents that are available on the World Wide Web and in corporate repositories increases, the ability to quickly and effectively search and view documents across language boundaries will continue to grow in importance. Cross-language information retrieval techniques allow searchers access to a wider range of material without requiring specialized knowledge of the content or the languages in the database. We present in this paper a cross-language information retrieval system based on a deep linguistic analysis of documents and queries and a statistical model which assigns a weight to each word in the database according to discriminating power. A comparison tool is used to evaluate all possible intersections between queries and documents and order documents by their relevance.

Keywords: Cross-language information retrieval, linguistic analysis, statistical model, bilingual dictionaries

1. Introduction

An information retrieval system accepts a query from a user and responds with a set of documents. 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 next one is slightly less likely and so on. Cross-lingual information retrieval aims to find relevant documents that are in a different language from that of the query [GREFENSTETTE 1998]. Bilingual query expansion ("reformulation") is often used to replace each query word by all its possible translations in target language documents.

This paper describes the cross-language information retrieval system developed at LIC2M [BESANCON & Al. 2003]. This system is designed to work on Arabic, Chinese, English, French, German, Italian, Spanish and Russian. It is based on developments and experiments undertaken during the first cross-language European project EMIR (European Multilingual Information Retrieval 1991-1994).

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

We present in section 2 the architecture of the cross-language information retrieval system, in particular, the linguistic and statistical approaches, the reformulation strategy and the indexing techniques. In section 3, the search engine of the cross-language information retrieval system is described, concentrating on document and query processing and strategies used for bilingual search and for merging the results. We present in section 4 the prototype of the search engine developed during the European project ALMA (Arabic Language Multilingual Application) and Section 5 discusses results obtained after submitting questions in Arabic, English and French.

2. Architecture of the LIC2M cross-lingual information retrieval system

The LIC2M cross-language retrieval system is a weighted Boolean search engine over syntactic structures produced by a linguistic analysis of the query and the documents. The system is composed of the following modules [SEMMAR & FLUHR 2004] (Figure 1):

- A linguistic analyzer, that processes both documents to be indexed and queries.
- A statistic analyzer, that computes concepts weights based on concepts database frequencies only for documents to be indexed.
- A reformulator, to expand queries during the search.
- A comparator, which computes semantic similarity between queries and indexed documents.
- An indexer to store indexed documents in a database.
- A search engine which searches the indexes for the closest documents to the expanded queries and merges the results obtained for each language.

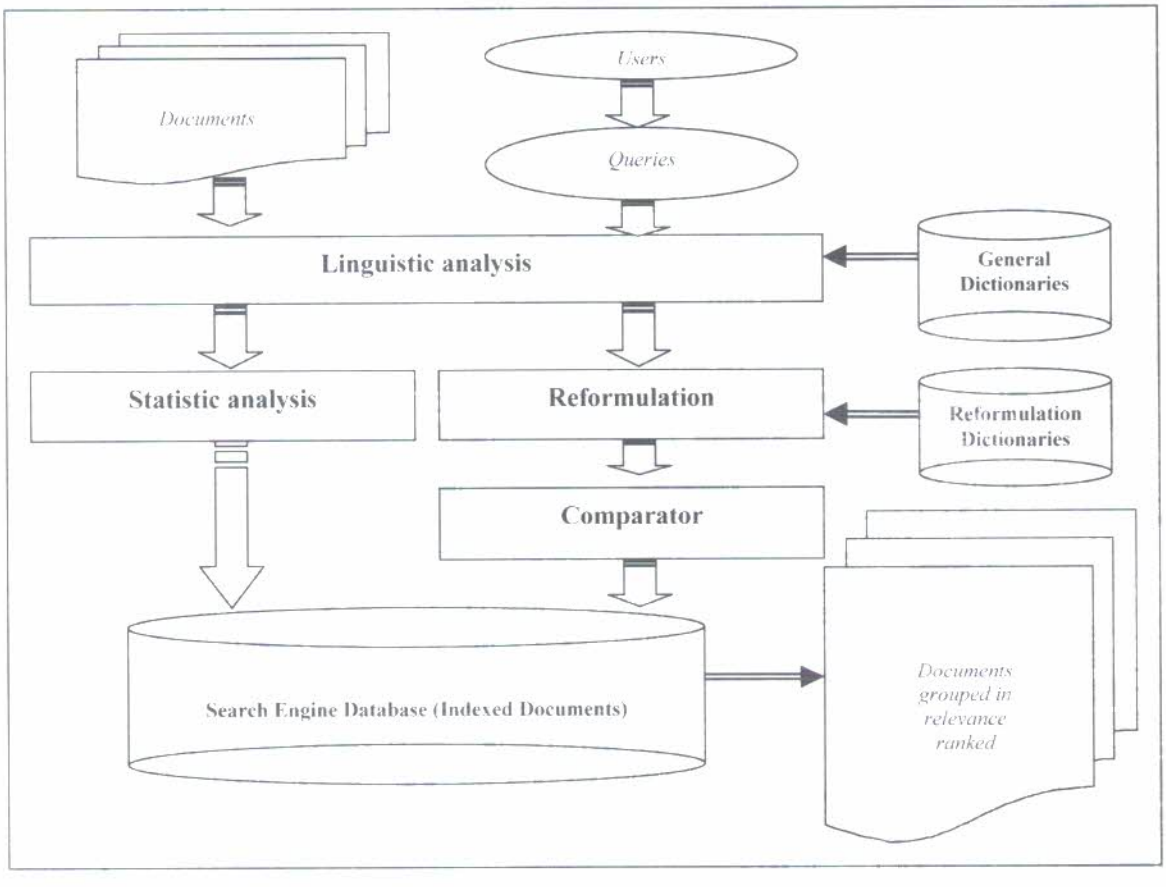

Documents
Users
Queries
Linguistic analysis
General Dictionaries
Statistic analysis
Reformulation
Reformulation Dictionaries
Comparator
Search Engine Database (Indexed Documents)
Documents grouped in relevance ranked

Figure 1: The LIC2M's cross-language retrieval system architecture.

2.1. Linguistic analysis

The linguistic analysis, a fundamental part of our cross-lingual information retrieval system, is composed of several linguistic resources (dictionaries, set of morpho-syntactic categories, etc.) and processing modules.

2.1.1. Linguistic resources

Each language has its proper linguistic resource components:

- A full form dictionary, containing for each word form its possible part-of-speech tags and linguistic features (gender, number, etc). For languages such as Arabic which present agglutination of articles, prepositions and conjunctions at the beginning of the word as well as pronouns at the ending of the word, we added two other dictionaries for proclitics and enclitics in order to split the input words into proclitics, simple forms and enclitics.
- A monolingual reformulation dictionary used in query expansion, for expanding original query words to other words expressing the same concepts (synonyms, hyponyms, etc.).
- Bilingual dictionaries used in cross-language querying.
- A set of rules for tokenizing words.
- A set of part-of-speech n-grams (bigrams and trigrams from hand-tagged corpora) that are used for part-of-speech tagging.
- A set of rules for shallow parsing of sentences, extracting compounds from the input text.
- A set of rules for the identification of named entities: gazetteers and contextual rules that use special triggers to identify named entities and their type.

Full form dictionary

All entries are accented, normalized and have linguistic properties like lemma and category. For Arabic, each entry is vowelled and associated to unvowelled versions as found in texts and queries. No linguistic properties are assigned to unvowelled entries which are naturally ambiguous, they are pointers towards the entry with vowels and its linguistic properties.

Unvowelled entries are produced by removing vowels and replacing certain letters: The vowels َ ِ ٍ ً ٌ ّ are removed, the characters أ آ إ are replaced by ا and the final characters ؤ ئ ي or ة are replaced by ءو ءى ى or ه.

The full form dictionary allows us to find all the vowelled entries corresponding to each word without vowels found in a corpus or a query. These vowelled entries correspond to the orthographic alternatives of the unvowelled surface word. For example, in the Arabic full form dictionary, the unvowelled word فتـــح has different linguistic properties according to its different vowellizations (Table 1):

Word	Morpho-syntactic tag	Lemma
فَتْحٌ	Noun, singular, masculine, in a nominative case	فتْح
فَتَحَ	Verb, past, 3 person, singular, masculine, active voice	فَتَحَ
فَتَّحَ	Verb, past, 3 person, singular, masculine, active voice	فَتَّحَ
فُتِحَ	Verb, past, 3 person, singular, masculine, passive voice	فَتَحَ

Table 1: Different linguistic properties and lemmas according to different vowels of the word.

This full form dictionary also allows us to find all the normalized entries corresponding to surface words found in texts. For example, once the word بـــريىء is found in the dictionary we have access to its possible normalizations: بَرِيىٌ بَرِيىُ بَرِيىً بَرِيىَ بَرِيىِ بَرِيى which are nouns or adjectives singular, masculine, in a nominative, accusative or genitive case. All these orthographic alternatives have بَرِيى as a lemma.

To produce the Arabic full form dictionary, we developed an inflector which automatically conjugates verbs and derives nouns [DEBILI & ZOUARI 1985] [ZOUARI 1989]. This tool produced 3,164,000 entries from 114,000 lemmas (nouns, adjectives and verbs). The final dictionary contains also closed lists like prepositions, pronouns, numbers, etc. This automatically generated dictionary is currently being manually corrected by native speakers.

For the other languages, we have acquired monolingual dictionnaries and modified them according to the structure of our full form dictionary.

Proclitics and enclitics dictionaries

Arabic proclitics and enclitics dictionaries have the same structure of the full form dictionary with vowelled and unvowelled versions of each clitic. They contain not only the individual proclitics and enclitics but all valid concatenations of these as well. No linguistic properties are assigned to concatenations of clitics. Each component of concatenated particles has its own linguistic properties. There are 77 and 65 entries respectively in each dictionary.

The table 2 contains some individual and compound entries from the dictionary of proclitics:

Individual proclitics	Compound proclitics
Prepositions: ب ,ك ,ل	وَلِلْ: Composed by the conjunction وَ, the preposition لِ and the definite article ال.
Conjunctions: فَ ,وَ	أَفَكَال: Composed by the particles أَ, the preposition فَ, the preposition كَ and the definite article ال.

Table 2: Individual and compound proclitics

The table 3 contains some entries from the dictionary of enclitics:

Enclitic	Enclitic morpho-syntactic tag
كُمَا ,كَ ,نِي ,هُمْ	Pronouns

Table 3: Some enclitics

Bilingual dictionaries

Bilingual dictionaries are used to reformulate the question in cross-language querying. We have purchased and modified transfer dictionaries for the following pairs of languages: Arabic-English, Arabic-French and English-French.

2.1.2. Processing modules

The processing modules are common for all the languages treated by the LIC2M analyzer with some variation for specific languages:

- Tokenizer.
- Morphological Analyzer.
- Recognizing Idiomatic Expressions.

- Part of Speech Tagger.
- Syntactic Analyzer.
- Recognizing Named Entities.
- Eliminating Empty Words.
- Normalizing Words.

The role of the **Tokenizer** is to split of characters strings into simple words. This tool takes into account context and segmentation rules to produce hopefully the best segmentation.

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 example, a word beginning with an uppercase letter obtains the tags of proper nouns.

At this point in the processing, an Arabic word that contains clitics will not have been found in the dictionary. We added a new processing step to the morphological analyzer for Arabic: a clitic stemmer [BUCKWALTER 2002].

The clitic stemmer proceeds as follows on unrecognized tokens:

- Several vowel form normalizations are performed (ً ُ ِ ّ are removed, أ إ آ are replaced by ا and final ؤ ئ ي or ة are replaced by وء ىء ى or ه).
- All clitic possibilities are computed by using proclitics and enclitics dictionaries.
- A radical, computed by removing these clitics, is checked against the full form lexicon. If it does not exist in the full form lexicon, re-write rules are applied, and the altered form is checked against the full form dictionary. For example, consider the token وهواهم and the included clitics (و, هم), the computed radical هوا does not exist in the full form lexicon but after applying one of the dozen re-write rules, the modified radical هوى is found in the dictionary and the input token is segmented into root and clitics as: وهواهم = و + هوى + هم.
- The compatibility of the morpho-syntactic tags of the three components (proclitic, radical, enclitic) is then checked. Only valid segmentations are kept and added into the word graph. Table 1 gives some examples of segmentations of words in the sentence من جانبها أكدت وزارة الداخلية العراقية.

Agglutinated word	Segmentations of the agglutinated word
ومن	ومن = و + من
جانبها	جانبها − جانب + ها

الداخلية	الداخلية = ال + داخلية
	الداخلية = [ا + ل] + داخلية

العراقية	العراقية = ال + عراقية
	العراقية = [ا + ل] + عراقية

المحافظات	المحافظات = ال + محافظات
	المحافظات = [ا + ل] + محافظات

للخطف	للخطف = [ل + ال] + خطف

الوزير	الوزير = ال + وزير
	الوزير = [ا + ل] − وزير

نفسه	نفسه = نفس + ه

Table 4: Segmentations of some agglutinated words.

For example, the agglutinated word الداخلية has two segmentations but only the segmentation: داخلية + ال = الداخلية will remain after POS tagging.

In table 4, we illustrate the final result of the morphological analysis of the sentence " ومن جانبها أكدت وزارة الداخلية العراقية تعرض 10 من حراس وزير شؤون المحافظات للخطف جنوب بغداد."

Word	Morpho-syntactic tags
و	Conjunction
من	Preposition
جانب	Verb, Noun, Adjective
ها	Pronoun
أكدت	Verb
وزارة	Noun
ال	Article
داخليـــة	Noun, Adjective
ال	Article
عراقية	Noun, Adjective
تعـــرض	Verb, Noun
10	Article, Noun, Adjective
من	Preposition
حراس	Noun, Adjective
وزيـــر	Noun, Adjective, Proper Noun
شـــؤون	Noun
ال	Article
محافظـــات	Noun, Adjective
ل	Preposition, Particle
ال	Article
خطف	Verb, Noun
جنـــوب	Preposition, Adjective, Noun
بغـــداد	Proper Noun
.	Punctuation

Table 5: Results of morphological analysis

The role of **Recognizing Idiomatic Expressions** 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 expressions can be non-contiguous, such as phrasal verbs: "*switch...on*"). 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. For Arabic, we have developed 482 contiguous expression rules. For example one of the developed rules recognizes in the text "كـــانون الثـــاني" as a whole and tags the expression as a being a month.

Part of Speech (POS) Tagging, one of the basic and indispensable tasks in the linguistic analysis, involves assigning to a word its disambiguated part of speech in the sentential context in which the word is used. Out of context, many words are ambiguous in their part of speech. For example, the word "تعـــرض" can feature as a noun or a verb. However when the word appears in the context of other words, the ambiguity is often reduced. For example, in "أكدت وزارة الداخلية العراقية تعرض 10 من حراس وزير شؤون المحافظات للخطف", the word "تعـــرض" can only be a noun.

Our linguistic analysis uses positional morpho-syntactic tags, meaning that the tag itself distinguishes which words can appear before or after another word. For example, for the Arabic language, there are pre-nominal and post-nominal adjectives. Pre-nominal adjectives can appear only before a noun and the post-nominal ones appear after a noun. Positional properties allow a very effective disambiguation, in general.

Our 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. If no continuous trigram full path is found, the algorithm 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.

Positional morpho-syntactic tags with adequate trigrams and bigrams matrices can completely disambiguate most of the words with a high precision. Trigrams and bigrams matrices are extracted from an annotated corpora of 13 200 words for Arabic, 239 000 words for English and 25 000 words for French.

For example, the Arabic sentence " ومن جانبها أكدت وزارة الداخلية العراقية تعرض 10 من حراس وزير شؤون المحافظات للخطف جنوب بغداد." is disambiguated by our tagger as follows (Table 6).

Word	Morpho-syntactic tag after POSTagging	Correct Morpho-syntactic tag
و	Conjunction	Conjunction
من	Preposition	Preposition
جانب	Noun	Noun
ها	Pronoun	Pronoun
أكدت	Verb	Verb
وزارة	Noun	Noun
ال	Article	Article
داخليـــة	Noun	Noun
ال	Article	Article
عراقية	Noun	**Adjective**
ضتعـــر	Noun	Noun
10	Noun	Noun
من	Preposition	Preposition
حراس	Noun	Noun
وزيــر	Proper Noun	**Noun**
شــؤون	Noun	Noun
ال	Article	Article
محافظــات	Adjective	**Noun**
ل	Preposition	Preposition
ال	Article	Article
خطف	Noun	Noun
جنــوب	Preposition	Preposition
بغــداد	Proper Noun	Proper Noun
.	.	Punctuation

Table 6: Results of disambiguation.

In table 7, we illustrate the results obtained with our POS Tagger for Arabic, English and French test corpora:

Language	Size of test corpora (words)	Accuracy (%)
Arabic	2 000	90.26
English	4 000	93.08
French	5 000	93.42

Table 7: Results of POS Tagger

We noticed that more than 50% of errors encountered in Arabic POS tagging result from confusing nouns and verbs. For example, the word تعــرض in تعرض الوزير لمحاولة إختطاف is alternately tagged as noun or an adjective. Other frequent errors result from confusing nouns with adjectives, these two categories are confusable.

The **Syntactic Analyzer** reveals syntactic structure of the analyzed sentence by using a set of syntactic rules. For example, the analyzer can split a sentence into nominal and verbal strings and recognize dependency relations (especially those within compounds). For the moment, for Arabic, we have developed only a small 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.

For example, the system links all these words because they are tagged as nouns and they are found in a same nominal chain (Figure 2):

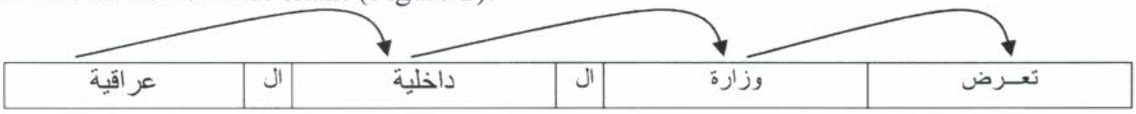


Figure 2: Dependency relations recognition.

The role **of Recognizing Named Entities** is to extract specific named entities such as names of persons, locations, organizations, products, events, dates and numbers. Specific named entities extraction is done by using a same method as that used to recognize idiomatic expressions. For example, in the Arabic sentence "ومن جانبها أكدت وزارة الداخلية العراقية تعرض 10 من حراس وزير شؤون المحافظات للخطف جنوب بغداد إلا أن الوزير نفسه لم يخطف.", two named entities are recognized (Table 8).

Named entity	Type
جنوب بغداد	Location
وزارة الداخلية العراقية	Organization

Table 8: Named entities recognition.

The location entity is recognized by the rule that stipulates: If we have in the text a word whose lemma is in this list (شَمَالِيّ شَمَال جَنُوبِي شَرْق غَرْب غَرْبِيّ شَرْقِي) followed by a Proper Noun, this sequence of word is tagged as a location.

Eliminating Empty Words consists in identifying words that should not be used as search criteria and eliminating them. These empty words are identified using only their parts of speech (such as prepositions, articles, punctuations and some adverbs).

Finally, words are **normalized** by their lemma. In the case that we have a set of synonymous lemmas, only one of these lemmas is taken as normalization.

It is often necessary to have the morpho-syntactic tag of the word to determine its meaning and sometimes its gender for nouns. For this reason, we also provide for each normalized word its morpho-syntactic tag.

2.2 Statistical analysis

The goal of the statistical processing in information retrieval is to be able to compare intersections between queries and documents, even if they contain different words. The statistical model is used to give the user a ranked list of documents, according to their relevance.

Our 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*".

By giving a weight to each word according to its "informativeness", we can weight the Boolean intersections between the words and phrases of query and those of the document.

2.3 Reformulation

In some cases, the linguistic processing and measure of the weight of the above intersection is not sufficient to establish a link between the query and the relevant documents. It is necessary to add a lexical semantic knowledge to the specified field in order to help (through the original query) the search engine to produce all the possible formulations of the same ideas that can occur in the various documents, candidate to be relevant. Query 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, hyponyms, etc.), or in different language (multilingual reformulation) [DEBILI et al. 1988].

2.4 Comparator

A comparison tool efficiently evaluates all possible intersections between query words and documents, and computes a relevance weight for each intersection.

3. Implementation of the Search Engine

Documents are processed to extract informative linguistic elements from the text parts. The processing includes a part-of-speech tagging of the words, their lemmatization and the extraction of compounds and named entities.

3.1 Query processing

All query processing is automatic. Each query is first processed through the linguistic analyzer corresponding to the query language. The result is a query composed of a list of the linguistic elements extracted from the analysis. These elements are called the concepts of the query. Each concept is reformulated into search terms in the language of the considered index, either using bilingual dictionaries or, in the case of monolingual search, using monolingual reformulation dictionaries.

3.2 Search and Merge strategies

For each language, the search engine retrieves for each term the documents containing the term. A concept profile is associated with each document, each component of which indicates the presence or absence of a query concept in the document (a concept is present in a document if at least one of its reformulated search terms is present). Retrieved documents sharing the same concepts profile are clustered together. This clustering allows for a straightforward merging strategy that takes into account original query concepts and the way they have been reformulated: since the concepts are in the original query language, the concept profiles associated with the clusters formed for different target languages are comparable since they are linked to the original language concepts, and the clusters even from different languages having the same profile are simply merged. To compute the relevance weight of each cluster, we compute a cross-lingual pseudo-idf weight of each concept.

4. The ALMA Prototype

The basic components of the ALMA prototype interact according to the following workflow:

- A Linguistic Analyzer: The output of this module is a list of pairs associating a word, which is a normalized form of either a simple word (its lemma), a compound noun or a named entity with its morpho-syntactic tag.
- An Indexer, which 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.
- A Query Processor, which reformulates queries on the basis of their linguistic analysis to suit the search (monolingual and multilingual reformulations): one reformulated query is built for each language of the document collection.
- 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 terms of the query (before reformulation) and their weights in order to score the documents.
- A Translation Engine, which is used to translate the retrieved documents (Systran Translation Engine is used via its web API form the Search Engine User Interface).

The user can enter a query in natural language and specify the language to be used. In the example of the Figure 3, the user entered the query "إدارة مـوارد الميـاه" and selected the *Arabic* as the language of the query.

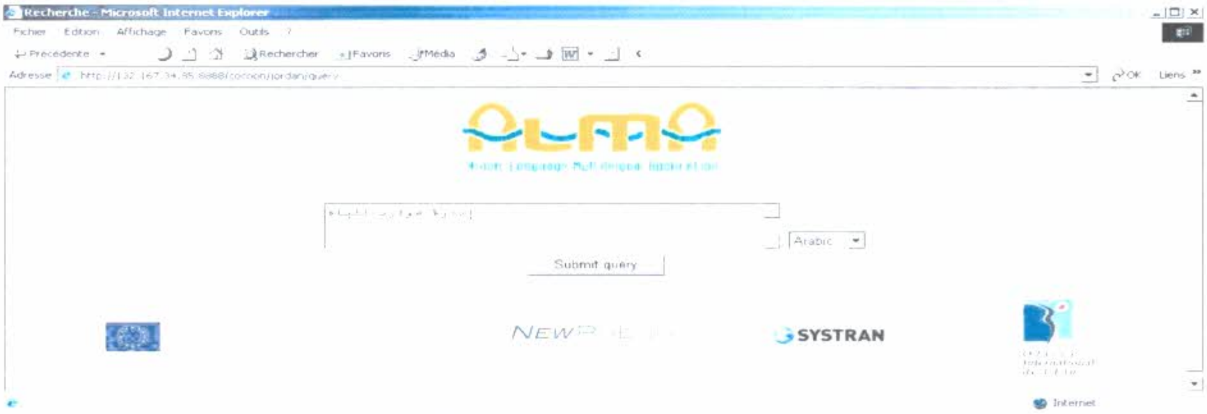
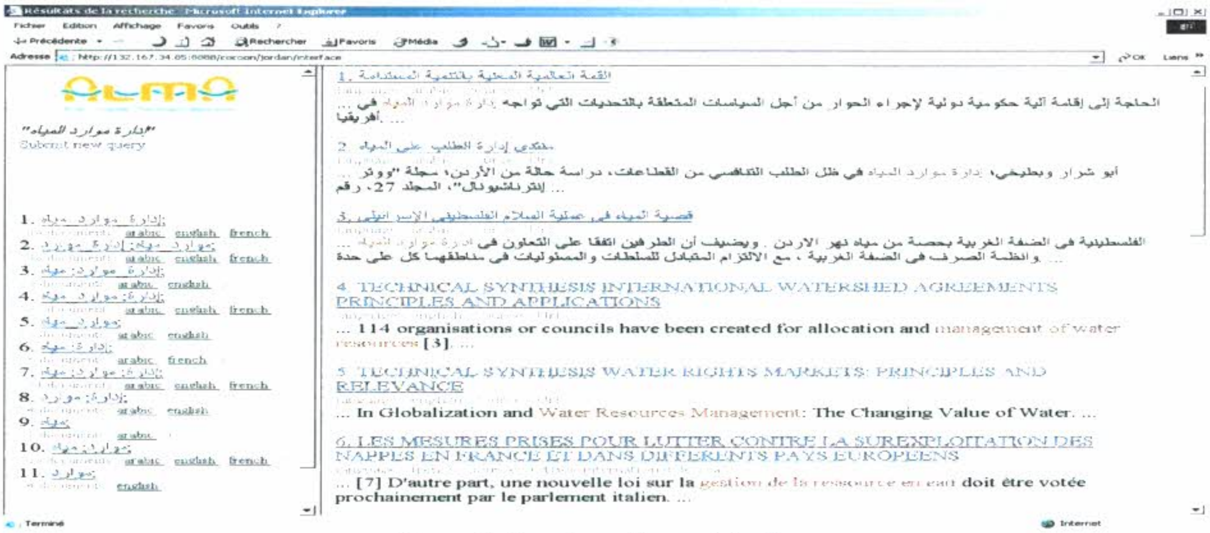


Figure 3: Queries user interface.

Relevant documents are grouped into classes characterized by the same set of concepts of the query. The classes constitute a discrete partition of the database (Figure 4).


Résultats de la recherche - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente Rechercher Favoris Média
Adresse http://132.167.34.85:8080/cocoon/jordan/interface
OK Liens
ALMA
Terminé
Internet

"إدارة موارد المياه"
Submit new query

1. إدارة_موارد_مياه; arabic english french
2. موارد_مياه; إدارة_موارد arabic english french
3. إدارة_موارد; مياه arabic english
4. إدارة; موارد_مياه arabic english french
5. موارد_مياه; arabic english
6. إدارة; مياه arabic french
7. إدارة; موارد; مياه arabic english french
8. إدارة; موارد arabic english
9. مياه; arabic
10. موارد; مياه arabic english french
11. موارد; english

1. القمة العالمية المحلية بالتنمية المستدامة
الحاجة إلى إقامة آلية حكومية دولية لإجراء الحوار من أجل السياسات المتعلقة بالتحديات التي تواجه إدارة موارد المياه في ... أفريقيا. ...

2. منتدى إدارة الطلب على المياه
أبو شرار وبطيخي، إدارة موارد المياه في ظل الطلب التنافسي من القطاعات، دراسة حالة من الأردن، مجلة "ووتر ... إنترناشيونال"، المجلد 27، رقم ...

3. قضية المياه في عملية السلام الفلسطيني الإسرائيلي
الفلسطينية في الضفة الغربية بحصة من مياه نهر الاردن . ويضيف أن الطرفين اتفقا على التعاون في ادارة موارد المياه ... وانظمة الصرف في الضفة الغربية ، مع الالتزام المتبادل للسلطات والمسئوليات في مناطقهما كل على حدة ...

4. TECHNICAL SYNTHESIS INTERNATIONAL WATERSHED AGREEMENTS PRINCIPLES AND APPLICATIONS
... 114 organisations or councils have been created for allocation and management of water resources [3]. ...

5. TECHNICAL SYNTHESIS WATER RIGHTS MARKETS: PRINCIPLES AND RELEVANCE
... In Globalization and Water Resources Management: The Changing Value of Water. ...

6. LES MESURES PRISES POUR LUTTER CONTRE LA SUREXPLOITATION DES NAPPES EN FRANCE ET DANS DIFFERENTS PAYS EUROPEENS
... [7] D'autre part, une nouvelle loi sur la gestion de la ressource en eau doit être votée prochainement par le parlement italien. ...

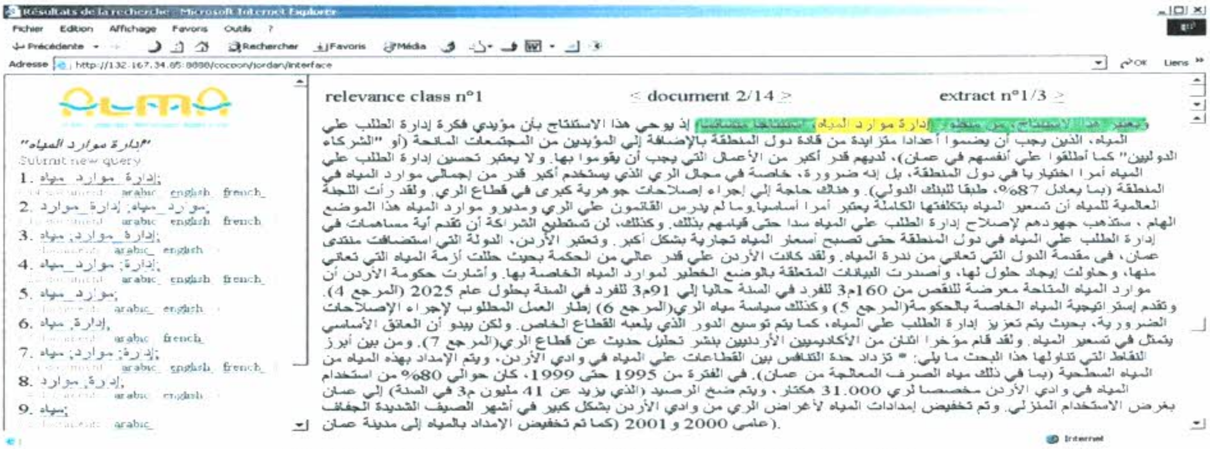
Figure 4: Search results user interface.

The table 9 illustrates some classes corresponding to the query "إدارة مـوارد الميـاه". The query term "إدارة_موارد_مياه" is a term composed of three words: مياه, موارد and إدارة. This compound word is computed by the syntactic analysis module.

Class	Query terms	Number of retrieved documents
1	إدارة_موارد_مياه	14
2	إدارة_موارد، موارد_مياه	18
3	مياه، إدارة_موارد	9

Table 9: First classes corresponding to the query "إدارة موارد المياه".

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


Résultats de la recherche - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente Rechercher Favoris Média
Adresse http://132.167.34.85:8080/cocoon/jordan/interface
OK Liens
ALMA
Internet

"إدارة موارد المياه"
Submit new query

1. إدارة_موارد_مياه; arabic english french
2. موارد_مياه; إدارة_موارد arabic english french
3. إدارة_موارد; مياه arabic english
4. إدارة; موارد_مياه arabic english french
5. موارد_مياه; arabic english
6. إدارة, مياه, arabic french
7. إدارة; موارد; مياه arabic english french
8. إدارة, موارد, arabic english
9. مياه; arabic

relevance class n°1 < document 2/14 > extract n°1/3 >

وتعتبر هذا الاستنتاج، من منظور إدارة موارد المياه، استنتاجا متناسبا إذ يوحي هذا الاستنتاج بأن مؤيدي فكرة إدارة الطلب على المياه، الذين يجب أن يضموا أعدادا متزايدة من قادة دول المنطقة بالإضافة إلى المؤيدين من المجتمعات المانحة (أو "الشركاء الدوليين" كما أطلقوا على أنفسهم في عمان)، لديهم قدر أكبر من الأعمال التي يجب أن يقوموا بها. ولا يعتبر تحسين إدارة الطلب على المياه أمرا اختياريا في دول المنطقة، بل إنه ضرورة، خاصة في مجال الري الذي يستخدم أكبر قدر من إجمالي موارد المياه في المنطقة (بما يعادل 87%، طبقا للبنك الدولي). وهناك حاجة إلى إجراء إصلاحات جوهرية كبرى في قطاع الري. ولقد رأت اللجنة العالمية للمياه أن تسعير المياه بتكلفتها الكاملة يعتبر أمرا أساسيا.وما لم يدرس القائمون على الري ومديرو موارد المياه هذا الموضوع الهام ، ستذهب جهودهم لإصلاح إدارة الطلب على المياه سدا حتى قيامهم بذلك. وكذلك، لن تستطيع الشراكة أن تقدم أية مساهمات في إدارة الطلب على المياه في دول المنطقة حتى تصبح أسعار المياه تجارية بشكل أكبر. وتعتبر الأردن، الدولة التي استضافت منتدى عمان، في مقدمة الدول التي تعاني من ندرة المياه. ولقد كانت الأردن على قدر عالي من الحكمة بحيث حللت أزمة المياه التي تعاني منها، وحاولت إيجاد حلول لها، وأصدرت البيانات المتعلقة بالوضع الخطير لموارد المياه الخاصة بها. وأشارت حكومة الأردن أن موارد المياه المتاحة معرضة للنقص من 160م3 للفرد في السنة حاليا إلى 91م3 للفرد في السنة بحلول عام 2025 (المرجع 4). وتقدم إستراتيجية المياه الخاصة بالحكومة(المرجع 5) وكذلك سياسة مياه الري(المرجع 6) إطار العمل المطلوب لإجراء الإصلاحات الضرورية، بحيث يتم تعزيز إدارة الطلب على المياه، كما يتم توسيع الدور الذي يلعبه القطاع الخاص. ولكن يبدو أن العائق الأساسي يتمثل في تسعير المياه. ولقد قام مؤخرا اثنان من الأكاديميين الأردنيين بنشر تحليل حديث عن قطاع الري(المرجع 7). ومن بين أبرز النقاط التي تناولها هذا البحث ما يلي: * تزداد حدة التنافس بين القطاعات على المياه في وادي الأردن، ويتم الإمداد بهذه المياه من المياه السطحية (بما في ذلك مياه الصرف المعالجة من عمان). في الفترة من 1995 حتى 1999، كان حوالي 80% من استخدام المياه في وادي الأردن مخصصا لري 31.000 هكتار ، ويتم ضخ الرصيد (الذي يزيد عن 41 مليون م3 في السنة) إلى عمان بغرض الاستخدام المنزلي. وتم تخفيض إمدادات المياه لأغراض الري من وادي الأردن بشكل كبير في أشهر الصيف الشديدة الجفاف (كما تم تخفيض الإمداد بالمياه إلى مدينة عمان عامي 2000 و 2001).

Figure 5: Highlighting query terms in retrieved documents.

5. Results

Our search engine has been tested on a multilingual corpora provided by partners of ALMA project. This base contains 50 non parrallel documents related to water, sustainable environment and tourism for each of the following languages: English, French and Arabic.We have found that results are better for English and French than for Arabic. A first analysis of the results suggests several possible adjustments:

- Monolingual reformulation introduces too many rare synonyms (or synonyms of rare senses of the words) that cause non-relevant documents to be retrieved. To improve the relevance of added terms the monolingual reformulation dictionaries must be revised.
- The importance of named entities was neglected in the queries we submitted. Giving a more importance to named entities, with added weight, should improve the results.

6. Conclusion

The results we obtained show that it is possible to do cross-language querying in a very robust way using linguistic, statistical approaches and bilingual dictionaries. However, inferior results obtained for Arabic Part of Speech tagging demonstrate the impact of the choice of the adequate morpho-syntactic tags and the size of the training corpora on the quality of disambiguation. To improve the quality of linguistic analysis for Arabic, we are currently working on finer morpho-syntactic tags and on the extraction of trigrams and bigrams sets from the Penn Arabic Treebank corpus (about 100,000 annotated words) [MAAMOURI & Al. 2004]. In addition, we are trying to improve the quality of the results by making the adjustments mentioned in Section 5.

References

[BESANCON & Al. 2003]
R. Besançon, Gaël de Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard and Hubert Naets, “The LIC2M’s CLEF 2003 system”, In Working Notes for the CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003.

[BUCKWALTER 2002]
T. Buckwalter, “Buckwalter Arabic Morphological Analyzer Version 1.0”, Linguistic Data Consortium, 2002.

[DEBILI & ZOUARI 1985]
F. Debili and L. Zouari, “Analyse morphologique de l'arabe écrit voyellé ou non fondée sur la construction automatique d'un dictionnaire arabe”, Cognitiva, Paris, France, 1985.

[DEBILI & Al. 1988]
F. Debili, C. Fluhr and P. Radasoa, “About reformulation in full text IRS”, Information processing and Management, England, 1988.

[GREFENSTETTE 1998]
G. Grefenstette, “Cross-language information retrieval”, Boston: Kluwer Academic Publishers, 1998.

[MAAMOURI & Al. 2004]
M. Maamouri, Ann Bies, Tim Buckwalter and Wigdan Mekki, “The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus”, NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 22-23 September 2004.

[SEMMAR & FLUHR 2004]
N. Semmar and C. Fluhr, "Multilingual Search Engine implementation", Final Technical report of ALMA project, EURO-MED programme, DG XIII, Commission of the European Union, Systran, France, July 2004.

[ZOUARI 1989]
L. Zouari, “Construction automatique d'un dictionnaire orienté vers l'analyse morpho-syntaxique de l'arabe, écrit voyellé ou non voyellé”, Thèse de doctorat, Université Paris XI, Paris, France, 1989.