# Using Stemming in Morphological Analysis to Improve Arabic Information Retrieval

Nasredine Semmar, Meriama Laib, Christian Fluhr

CEA – Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue
{nasredine.semmar; meriama.laib; christian.fluhr}@cea.fr

## Résumé

La recherche d'information consiste à trouver les documents pertinents parmi un ensemble de documents en réponse à une requête de l'utilisateur. Ces documents sont triés par ordre de pertinence. Le but du traitement automatique du langage naturel dans la recherche d'information est de transformer les mots potentiellement ambigus de la requête et des documents en représentations internes non ambiguës sur lesquelles s'effectuera l'appariement. Cette transformation est généralement réalisée à l'aide de plusieurs niveaux d'analyse linguistique (morphologique, syntaxique, etc.). Cet article présente l'analyseur linguistique de l'arabe du moteur de recherche crosslingue du LIC2M. Nous allons nous concentrer sur l'analyseur morphologique et plus particulièrement sur le module de segmentation qui permet de découper les mots agglutinés en proclitiques, formes simples et enclitiques. Nous allons démontrer qu'une bonne segmentation améliore la précision et le rappel du moteur de recherche.

**Mots-clés** : analyse morphologique, désambiguïseur morpho-syntaxique, analyse syntaxique, entités nommées, découpage, recherche d'information crosslingue.

## Abstract

Information retrieval (IR) consists in finding all relevant documents for a user query in a collection of documents. These documents are ordered by the probability of being relevant to the user's query. The highest ranked document is considered to be the most likely relevant document. Natural Language Processing (NLP) for IR aims to transform the potentially ambiguous words of queries and documents into unambiguous internal representations on which matching and retrieval can take place. This transformation is generally achieved by several levels of linguistic analysis, morphological, syntactic and so forth. In this paper, we present the Arabic linguistic analyzer used in the LIC2M cross-lingual search engine. We focus on the morphological analyzer and particularly the clitic stemmer which segments the input words into proclitics, simple forms and enclitics. We demonstrate that stemming improves search engine recall and precision.

**Keywords**: morphological analysis, part-of-speech tagging, syntactic analysis, named entities, stemming, cross-lingual information retrieval.

# 1. Introduction

Most Arabic words are composed of a basic form that is called radical and many attached affixes (proclitics and enclitics). Proclitics are attached in the beginning of the word and can be articles, prepositions and conjunctions and enclitics are attached at the end of the word and are in general pronouns. This abundance of forms has an impact on the precision of information retrieval applications results (the form of the word in a query is different from the forms found in documents). To resolve this vocabulary mismatch problem, we use stemming during the linguistic analysis of queries and documents. The stemmer splits agglutinated words into proclitics, simple forms and enclitics and allows the search engine to index only simple forms.

We present in section 2, the main components of our linguistic analyzer, in particular, the morphological analyzer and the linguistic resources. In section 3, the prototype of the cross-lingual search engine developed during the European project ALMA (Arabic Language Multilingual Application) is described. We present in section 4 the experimental results obtained with our cross-lingual search engine and we discuss the effect of stemming for both mono-lingual and cross-lingual information retrieval. In section 5, we conclude our study and we present our future work.

## 2. Linguistic analysis

Linguistic analysis is a fundamental part in the LIC2M cross-language information retrieval, since it determines the syntactic structures of the queries and documents sentences (Grefenstette *et al.*, 2005) on which matching and retrieval can take place. The LIC2M linguistic analysis is composed of the following modules:

1. Morphological analysis:
   - Simple word lookup to search for words in a full form lexicon.
   - Orthographical alternative lookup to look for differently accented forms, alternative hyphenisation, concatenated words and abbreviation recognition.
   - Clitic stemmer to split unknown input Arabic words into proclitics, simple forms and enclitics.
   - Idiomatic expressions recognizer to detect idiomatic expressions and consider them as single words in the word graph.
   - Unknown word analysis.
2. Part-of-Speech Tagging reduces the number of possible morpho-syntactic tags of the input words using language models from a hand-tagged corpus.
3. Syntactic analysis splits input words into nominal and verbal chain and recognizes dependency relations.
4. Named entity recognizer identifies names of locations, organizations, persons, etc. by using name triggers.

### 2.1. Morphological analysis

After Tokenization which consists in separating the input stream into a graph of words by taking into account context and segmentation rules, the morphological analyzer proceeds as follows:

The Simple word lookup searches for words in a full form lexicon which is stored as a finite-state automata. Entries of the full form lexicon are associated to their lemmas, morpho-syntactic tags, linguistic properties (gender, number, etc), normalized and accented/non accented forms. No morpho-syntactic tags or linguistic properties are assigned to non accented forms which are naturally ambiguous. These entries have pointers towards the corresponding accented words and their linguistic properties. For languages such as French and English, a non accented form can be a word without any accent or an orthographic alternative. For Arabic, each entry is vowelled and associated to unvowelled versions.

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

| Word | Morpho-syntactic tag | Lemma |
|---|---|---|
| فَتْحٌ (victory) | Noun, singular, masculine, in a nominative case | فَتْح |
| فَتَحَ (to open) | Verb, past, 3 person, singular, masculine, active 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 ءىيرب (innocent) is found in the dictionary we have access to its possible vocalizations: بَرِيئٌ بَرِيئُ بَرِيئً بَرِيئ يىَبَر بَرِيئِ بَرِيئ which are nouns or adjectives singular, masculine, in a nominative, accusative or genitive case. All these orthographic alternatives have بَرِيئ as a lemma.

To produce the Arabic full form dictionary, we developed an inflector which automatically conjugates verbs and derives nouns (Debili and 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 dictionaries and modified them according to the structure of our full form dictionary.

The Orthographical alternative lookup looks for differently accented forms, alternative hyphenisation, concatenated words, abbreviation recognition, which might alter the original non-cyclic word graph by adding alternative paths. At this point in the processing, a word that contains clitics will not be in the dictionary since we had decided not to include word forms including clitics (Attia, 1999). We added a new processing step for Arabic: a clitic stemmer (Larkey *et al.*, 2002). This stemmer uses a full form dictionary, a proclitic dictionary and an enclitic dictionary.

The Clitic stemmer proceeds as follows on tokens unrecognized after orthographical alternative lookup:

1. 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 ءو ءى ى or ه.
2. All clitic possibilities are computed by using proclitics and enclitics dictionaries.
3. 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 (such as those described in Darwish, 2002) are applied, and the altered form is checked against the full form dictionary. For example, consider the token « بكرته » (with its ball) and the included clitics ب and ه, the computed radical[1] كرت does not exist in the full form lexicon but after applying one of the re-write rules, the modified radical « ةرك » (ball) is found in the dictionary and the input token is segmented into root and clitics as: ه + كرة + ب = بكرته (with + its + ball).

[1] The word "كرت" does not exist in our full form dictionary, its use in Arabic corresponds to the transleteration of the french word "carte".

4. 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 2 gives some examples of segmentations[2] of words in the sentence « المياه الصالحة للشرب » (drinking water).

| Agglutinated word | Segmentations of the agglutinated word |
|---|---|
| المياه | المياه = ا + لميا + ه<br>المياه = ال + مياه<br>المياه = [ا + ل] + مياه<br>المياه = الميا + ه |
| الصالحة | الصالحة = ال + صالحة<br>الصالحة = [ا + ل] + صالحة |
| للشرب | للشرب = [ل + ال] + شرب |

*Table 2. Segmentations of some agglutinated words*

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 proclitics as well. No linguistic properties are assigned to concatenations of clitics. Each component of concatenated proclitics has its own linguistic properties. There are 77 and 65 entries respectively in each dictionary. Table 3 contains some individual and compound entries from the dictionary of proclitics and the table 4 contains some entries from the dictionary of enclitics:

| Proclitics | Proclitics morpho-syntactic tags |
|---|---|
| لِ كَ بِ | Prepositions |
| وَلِال | Composed by the conjunction وَ, the preposition لِ and the definite article ال. |

*Table 3. Individual and compound proclitics*

| Enclitic | Enclitic morpho-syntactic tags |
|---|---|
| هُمْ نِي كَ كُمَا | Pronouns |

*Table 4. Some enclitics*

The Idiomatic expressions recognizer is used to detect idiomatic expressions and consider them as single words in the word graph. This operation[3] is performed after clitic separation

[2] For example, the agglutinated word المياه has 4 segmentations but only the segmentation: المياه = ال + مياه will remain after POS tagging.

using rules associated with trigger words for each expression. Once a trigger is found, its left and right lexical contexts in the rule are then tested. The trigger must be an entry in the full form lexicon, but can be represented as either a surface form or a lemma form combined with its morpho-syntactic tag. Because Arabic lexicon entries are vowelled and input texts may be partially vowelled or unvowelled, we used only lemma forms to describe Arabic idiomatic expressions rules. We developed 482 contiguous idiomatic vowelled expression rules. For example one of the developed rules recognizes in the text « كانون الثاني » (January) as a whole and tags the expression as being a month.

The Unknown words processing module assigns to the nodes not yet recognized default linguistic values based on features recognized during tokenization (*e.g.* presence of Arabic characters, Latin characters, numbers, special characters, etc.).

## 2.2. Part of speech tagging

One of the basic and indispensable tasks in linguistic analysis consists in 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 « خطف » (kidnapping) 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 the sentence « أكدت الشرطة خطف صحفي » (security forces confirmed the kidnapping of a journalist) the word « خطف » 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.

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. 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 (Semmar *et al.*, 2005). Table 5 illustrates 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 5. Performance of the POS Tagger*

## 2.3. Syntactic analysis

After part-of-speech tagging, a syntactic analyzer is used to split word graph into nominal and verbal chain and recognize dependency relations by using a set of syntactic rules. For Arabic,

[3] An idiom is a (possibly non-contiguous) sequence of known words that act as a single unit. Once an idiomatic expression is recognized the individual word nodes are joined into one node in the word graph.

the only dependency relations that we extract for the moment are relations between nominal elements (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 « إستعمال الطاقة الذرية » (use of nuclear energy) the system links words as following (figure 1) because "إستعمال" (use) and "طاقة" (energy) are tagged as nouns and "ذرية" (nuclear) as an adjective.

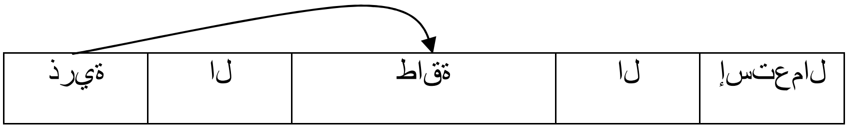


*Figure 1. Dependency relations recognition*

These relations are used to compute three compound words that are normalized as إستعمال_طاقة_ذرية (use_nuclear_energy), طاقة_ذرية (nuclear_energy), and إستعمال_طاقة (use_energy).

### 2.4. Named Entities Recognition

The next step in the LIC2M linguistic analyzer, after syntactic analysis, is named entity recognition (Abuleil and 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. In Arabic, the input text is usually unvowelled or partially vowelled (the same named entity has different surface forms). To resolve this problem, we stored all the vowelled forms of the entity in the full form lexicon and we linked the unvowelled entity to these vowelled forms. The simple word lookup provides the different vowelled forms of the entity which are used by the named entities recognition rules.

For example, in the Arabic sentence « توزيع المياه الصالحة للشرب جنوب العراق » (supply of drinking water in the south of Irak), « جنوب العراق » (south of Irak) is recognized as a named entity corresponding to a « Location ». This named entity is tagged as a « Location » because the word « العراق » (Irak) is a proper noun and it follows the lemma « جنوب » (south).

## 3. The ALMA search engine prototype

Most of Arabic commercial search engines are full form based retrieval. They do not use stemming to analyze query words and to index documents (Abdelali *et al.*, 2004). For example, querying with the Arabic words اقتصاد or إقتصاد or الإقتصاد do not return the same results. These results are acceptable for web users but they are not relevant in electronic document management applications.

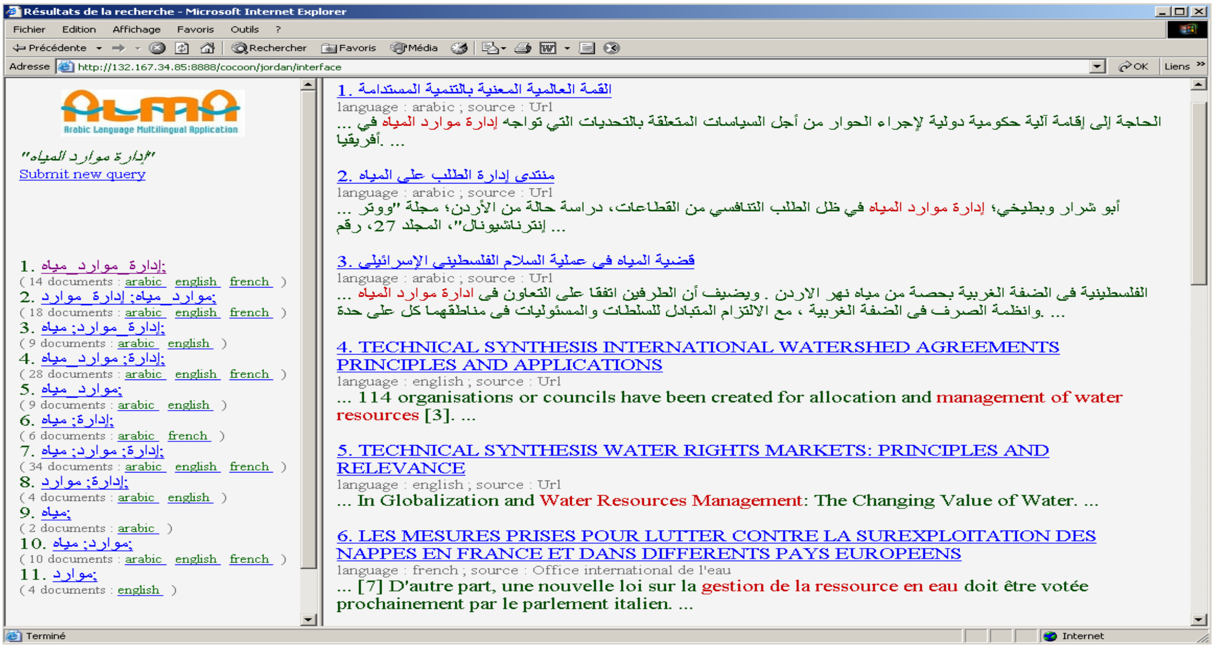
The goal of the European project ALMA was to develop a web solution based on a multilingual information retrieval system and an automatic translation tool. This solution was expected to use a deep linguistic analysis to process queries and documents to be indexed. To achieve this goal, the ALMA partners were involved in the main following tasks:

- Collection of a consistent corpus of documents in English, French and Arabic, on the domains of sustainable development, water and eco-tourism.
- Implementation of a search engine and an automatic translation tool to provide access to multilingual information.

The ALMA search engine prototype is based on the LIC2M cross-language information retrieval system and is visible online at a third party site: http://alma.oieau.fr (Semmar and Fluhr, 2004). It 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 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, hyponyms, etc.) or in different language.
- A comparator, which computes intersections between queries and documents and provides a relevance weight for each intersection.
- 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.

The user of the ALMA search engine prototype can enter a query in natural language and specify the language to be used. In the example of the figure 2, the user entered the query “هايملا دراوم ةرادإ” (water resources management) in Arabic and relevant retrieved documents are grouped into classes. Each class is characterized by the same set of concepts. For example, the first class « إدارة_موارد_مياه » is characterized by a term composed of three words: مياه (water), موارد (resources) and إدارة (management). This compound word is computed by the syntactic analysis module.


Résultats de la recherche - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente Rechercher Favoris Média
Adresse http://132.167.34.85:8888/cocoon/jordan/interface
OK Liens
Arabic Language Multilingual Application
Terminé
Internet

"إدارة موارد المياه"
Submit new query

1. إدارة_موارد_مياه;
( 14 documents : arabic english french )
2. موارد_مياه; إدارة_موارد
( 18 documents : arabic english french )
3. إدارة_موارد; مياه
( 9 documents : arabic english )
4. إدارة; موارد_مياه
( 28 documents : arabic english french )
5. موارد_مياه;
( 9 documents : arabic english )
6. إدارة; مياه
( 6 documents : arabic french )
7. إدارة; موارد; مياه
( 34 documents : arabic english french )
8. إدارة; موارد
( 4 documents : arabic english )
9. مياه;
( 2 documents : arabic )
10. موارد; مياه
( 10 documents : arabic english french )
11. موارد;
( 4 documents : english )

1. القمة العالمية المعنية بالتنمية المستدامة
language : arabic ; source : Url
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
language : arabic ; source : Url
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 فى عملية السلام الفلسطينى الإسرائيلى
language : arabic ; source : Url
الفلسطينية فى الضفة الغربية بحصة من مياه نهر الاردن . ويضيف أن الطرفين اتفقا على التعاون فى ادارة موارد المياه ...
وانظمة الصرف فى الضفة الغربية ، مع الالتزام المتبادل للسلطات والمسئوليات فى مناطقهما كل على حدة. ...

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

5. TECHNICAL SYNTHESIS WATER RIGHTS MARKETS: PRINCIPLES AND RELEVANCE
language : english ; source : Url
... 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
language : french ; source : Office international de l'eau
... [7] D'autre part, une nouvelle loi sur la gestion de la ressource en eau doit être votée prochainement par le parlement italien. ...

*Figure 2. Search results of the query* “هايملا دراوم ةرادإ”

## 4. Results and discussion

We submitted to the LIC2M search engine prototype two runs for both mono-lingual and cross-lingual retrieval: one with activating the stemmer during the morphological analysis and the other without activating the stemmer. The run consists in submitting questions in Arabic and to validating the relevance of retrieved documents. We checked manually the relevance of matching of 100 questions against 50 non-parallel documents for each language (Arabic, English and French). Each Arabic document is relevant for 2 questions (the first question is extracted directly from the text of the document and the second is related to the document but is not a part of it). The 50 non-parallel documents were provided by the ALMA partners and are related to sustainable development, water and eco-tourism.

Figure 3 shows for each run the precision-recall curve for cross-lingual retrieval and figure 4 shows the precision-recall curve of the two runs for mono-lingual retrieval.

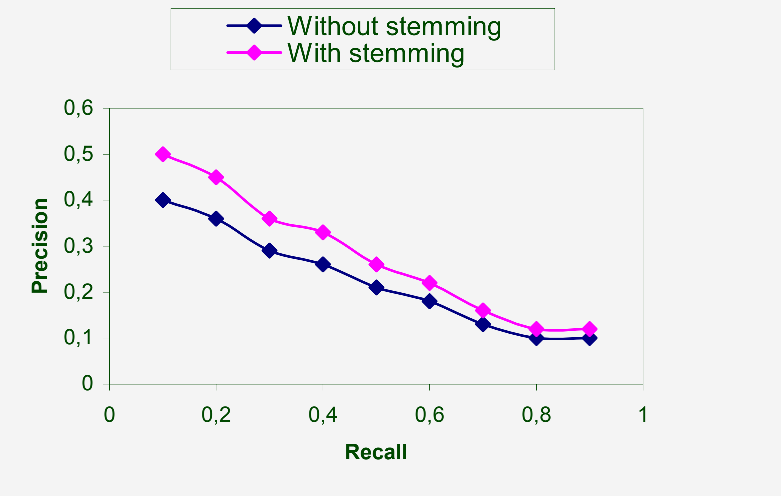


*Figure 3. Results of the two runs for cross-lingual retrieval*

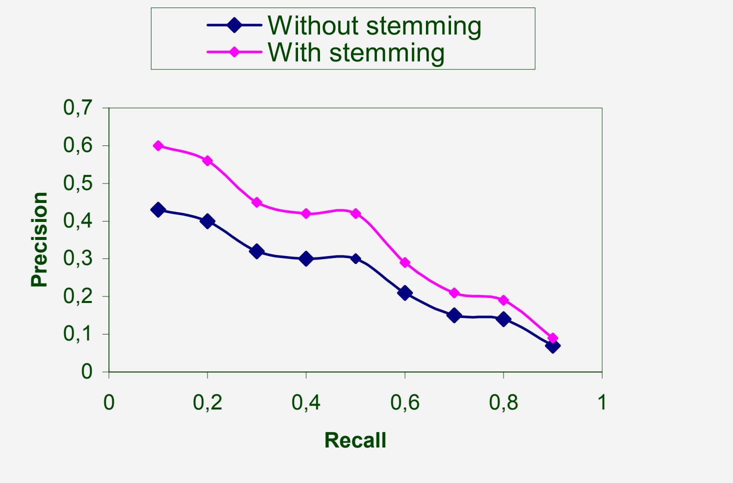


*Figure 4. Results of the two runs for mono-lingual retrieval*

The results of figures 3 and 4 show increase in precision around 41 % for mono-lingual retrieval and 27 % for cross-lingual retrieval if the stemmer is activated.

It appears, in general, that results are better for mono-lingual retrieval than cross-lingual retrieval. This comes from the fact that our search engine uses not only lemmas but also parts-of-speech of query terms during lookup in the bilingual dictionaries.

As expected, stemming improves precision in both mono-lingual and cross-lingual retrieval. This is because the stemmer splits agglutinated Arabic words into proclitics, simple forms and enclitics and the search engine considers proclitics and enclitics as empty words and does not index them in the database.

Concerning increase in recall, the linguistic analyzer normalizes the simple form of the agglutinated Arabic word (verb in infinitive, noun in singular form, etc.) and the search engine indexes all the derivative forms in the database.

## 5. Conclusion and future work

Our experiments showed that stemming significantly contributes to improve Arabic information retrieval. These experiments confirmed results obtained by Larkey and al. (Larkey *et al.*, 2002) with their light stemmer which removes only stop words, definite articles, conjunctions and preposition from the beginning of the words and small number of suffixes from the end of words. On the other hand, our experiments also demonstrated the impact of bilingual dictionaries on the effectiveness of cross-lingual retrieval. In order to confirm our approach, we are currently working on a generic version of our stemmer to process other languages which have a similar phenomenon (Spanish, Finnish, Hungarian, etc.) and we are going to acquire TREC 2002 Arabic corpus to experiment our search engine on a large database.

## References

ABDELALI A., COWIE J., SOLIMAN H.S. (2004). "Arabic Information Retrieval Perspectives". In *Actes de TALN 2004*.

ABULEIL S., EVENS M. (2004). "Named Entity Recognition and Classification for Text in Arabic". In *Actes du IASSE 2004*: 89-94.

ATTIA M. (1999). *A large-Scale Computational Processor of Arabic Morphology and Applications.* M.S. thesis in Computer Engineering, Cairo University: 28-32.

DARWISH K. (2002). "Building a Shallow Arabic Morphological Analyzer in One Day". In *Proceedings of ACL-2005* : 47-54.

DEBILI F., ZOUARI L. (1985). "Analyse morphologique de l'arabe écrit voyellé ou non fondée sur la construction automatique d'un dictionnaire arabe". In *Actes du Cognitiva-1985*.

GREFENSTETTE G., SEMMAR N., ELKATEB-GARA F. (2005). "Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information". In *Processing and Information Retrieval Applications. Proceedings of ACL-200*: 31-38.

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*: 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 *Actes du Deuxième Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la Langue*: 114–125.

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

ZOUARI L. (1989). *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.