Arabic to French Sentence Alignment: Exploration of A Cross-language Information Retrieval Approach

Nasredine Semmar
CEA, LIST
Laboratoire d'ingénierie de la connaissance multimédia multilingue
18 route du Panorama
BP6, FONTENAY AUX ROSES, F-92265 France
nasredine.semmar@cea.fr

Christian Fluhr
CEA, LIST
Service Réalite virtuelle, Cognitique et Interfaces
18 route du Panorama
BP6, FONTENAY AUX ROSES, F-92265 France
christian.fluhr@cea.fr

Abstract

Sentence alignment consists in estimating which sentence or sentences in the source language correspond with which sentence or sentences in a target language. We present in this paper a new approach to aligning sentences from a parallel corpus based on a cross-language information retrieval system. This approach consists in building a database of sentences of the target text and considering each sentence of the source text as a "query" to that database. The cross-language information retrieval system is a weighted Boolean search engine based on a deep linguistic analysis of the query and the documents to be indexed. This system is composed of a multilingual linguistic analyzer, a statistical analyzer, a reformulator, a comparator and a search engine. The multilingual linguistic analyzer 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. The statistical analyzer computes for documents to be indexed concept weights based on concept database frequencies. The comparator computes intersections between queries and documents and provides a relevance weight for each intersection. Before this comparison, the reformulator expands queries during the search. The expansion is used to infer from the original query words other words expressing the same concepts. The search engine 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 sentence aligner has been evaluated on the MD corpus of the ARCADE II project which is composed of news articles from the French newspaper "Le Monde Diplomatique". The part of the corpus used in evaluation consists of the same subset of sentences in Arabic and French. Arabic sentences are aligned to their French counterparts. Results showed that alignment has correct precision and recall even when the corpus is not completely parallel (changes in sentence order or missing sentences).

1 Introduction

Sentence alignment consists in mapping sentences of the source language with their translations in the target language. Automatic sentence alignment approaches face two kinds of difficulties: robustness and accuracy. A number of automatic sentence alignment techniques have been proposed (Kay and Röscheisen, 1993; Gale and Church, 1991; Brown et al., 1991; Debili and Samouda, 1992; Papageorgiou et al., 1994; Gaussier, 1995; Melamed, 1996; Fluhr et al., 2000).

Proceedings of the 5th Workshop on Important Unresolved Matters, pages 73–80,
Prague, Czech Republic, June 2007.

The method proposed in (Kay and Röscheisen, 1993) is based on the assumption that in order for the sentences in a translation to correspond, the words in them must correspond. In other words, all necessary information (and in particular, lexical mapping) is derived from the to-be-aligned texts themselves.

In (Gale and Church, 1991) and (Brown et al., 1991), the authors start from the fact that the length of a source text sentence is highly correlated with the length of its target text translation: short sentences tend to have short translations, and long sentences tend to have long translations.

The method proposed in (Debili and Sammouda, 1992) is based on the preliminary alignment of words using a conventional bilingual lexicon and the method described in (Papageorgiou et al., 1994) added grammatical labeling based on the assumption that the same parts of speech tend to be employed in the translation.

In this paper, we present a sentence aligner which is based on a cross-language information retrieval approach and combines different information sources (bilingual lexicon, sentence length and sentence position). This sentence aligner was first developed for aligning French-English parallel text. It is now ported to Arabic-French and Arabic-English language pairs.

We present in section 2 the main components of the cross-language search engine, in particular, we will focus on the linguistic processing. In section 3, the prototype of our sentence aligner is described. We discuss in section 4 results obtained after aligning sentences of the MD (Monde Diplomatique) corpus of the ARCADE II project. Section 5 concludes our study and presents our future work.

2 The Cross-language Search Engine

Information retrieval consists to find 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. Cross-language information retrieval consists in providing a query in one language and searching documents in different languages (Grefenstette, 1998). The cross-lingual search engine is a weighted Boolean search engine based on a deep linguistic analysis of the query and the documents to be indexed (Besançon et al., 2003). It is composed of a linguistic analyzer, a statistical analyzer, a reformulator and a comparator (Figure 1):

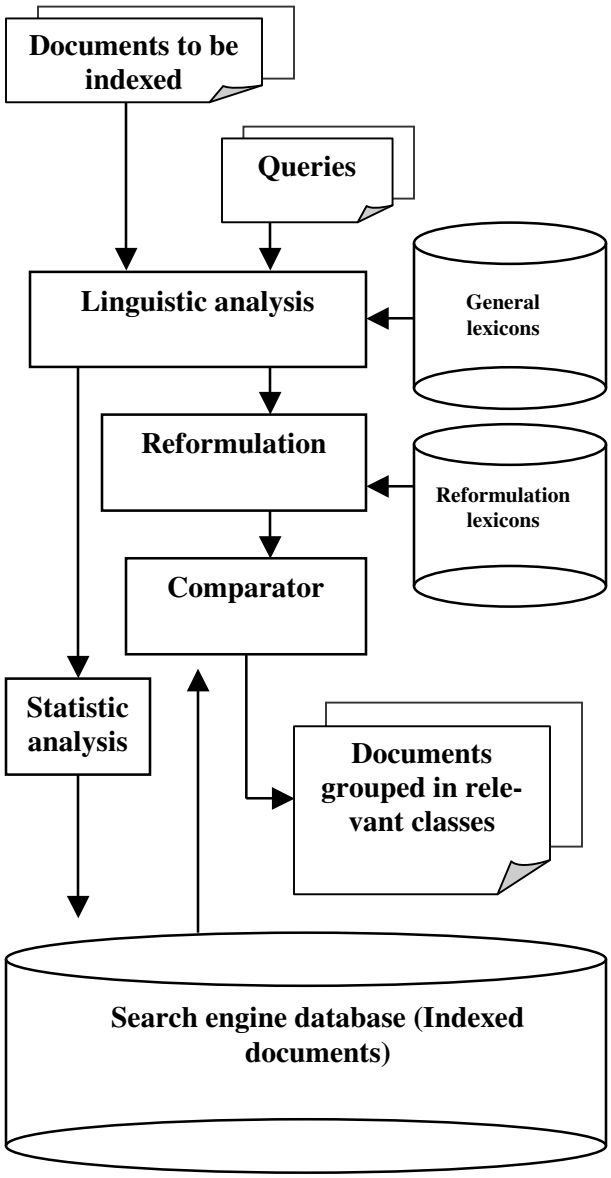

Figure 1. The cross-language search engine

2.1 Linguistic Analysis

The linguistic analyzer produces a set of normalized lemmas, a set of named entities and a set of nominal compounds. It is composed of several linguistic resources and processing modules.

Each language has its proper linguistic resources which are generally composed of:

- 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 presents 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.

The processing modules are common for all the languages with some variations for some specific languages:

- A Tokenizer which 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.
- A Morphological analyzer which searches each word in a general dictionary (Debili and Zouari, 1985). 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. If the simple form computed by the clitic stemmer does not exist in the general dictionary, re-write rules are applied (Darwish, 2002). For example, consider the token "بكرتهم" (with their ballon) and the included clitics "ب" (with) and "هم" (their), the computed simple form "كرت" does not exist in the general dictionary but after applying one of the dozen re-write rules, the modified simple form "كرة" (ballon) is found in the general dictionary and the input token is segmented as: بكـــرتهم = ب + كرة + هم.
- An Idiomatic Expressions recognizer which detects idiomatic expressions and considers 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 "البَيْت الأبْيَض" in Arabic. Non-contiguous expressions such as phrasal verbs in English: "switch…on" or "tomber vaguement dans les pommes" in French are recognized too.
- A Part-Of-Speech (POS) tagger which 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 (Grefenstette et al., 2005). They are extracted from a hand-tagged corpora of 13 200 words for Arabic 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. The accuracy of the part-of-speech tagger is around 91% for Arabic and 94% for French.
- A Syntactic analyzer which 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 "توزيع المياه" (water supply), the syntactic analyzer considers this nominal chain as a compound word (توزيع_مياه) composed of the words "توزيع" (supply) and "مياه" (water).

- A Named Entity recognizer which uses name triggers (e.g., President, lake, corporation, etc.) to identify named entities (Abuleil and Evens, 2004). For example, the expression "الأوَّل مِن شَهْر مَارس" (The first of March) is recognized as a date and the expression "الشَّرْق الأوْسَط" (The Middle East) is recognized 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 the word has a set of synonymous lemmas, only one of these lemmas is taken as a normalization. Each normalized word is associated with its morpho-syntactic tag.

2.2 Statistical Analysis

The role of the statistical analysis is to attribute a weight to each word or a compound word according to the information the word or the compound word provides in choosing the document relevant to a query. This weight is computed by an idf formula (Salton and McGill, 1983). 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 compute the semantic intersection between query and documents containing different words. A similarity value is associated with each semantic intersection. This value corresponds to the sum of the weights of words present in the documents. The search engine groups documents into classes (semantic intersections) characterized by the same set of words. These classes constitute a discrete partition of the indexed documents. For example, the search engine returns 12 classes for the query "إدارة موارد المياه" (water resources management) (Table 1).

Class	Query terms
1	إدارة_موارد_مياه
2	موارد_مياه ,إدارة_موارد
3	مياه ,إدارة_وارد
4	إدارة ,موارد_مياه
5	إدارة_موارد
6	موارد_مياه
7	إدارة ,موارد ,مياه
8	إدارة ,مياه
9	إدارة ,موارد
10	موارد ,مياه
11	مياه
12	موارد

Table 1. Relevant classes returned by the search engine for the query "إدارة موارد المياه"

The query term "إدارة_موارد_مياه" is a compound word composed of three words: "إدارة" (management), "موارد" (resources) and "مياه" (water). This compound word is computed by the syntactic analyzer.

2.3 Query Reformulation

The role of query reformulation is to infer new words from the original query words according to a lexical semantic knowledge. The reformulation can be used to increase the quality of the retrieval in a monolingual interrogation. 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. Reformulation rules can be applied to all instances of a word or to a word only when it is playing a specific part-of-speech. Semantic relations can also be selected: translations, synonyms, word derived from the same root, etc. The cross-language search engine has a monolingual reformulation for French and two bilingual reformulations for Arabic-French and French-Arabic language pairs.

2.4 Query and Documents Comparison

The 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 search engine uses a comparison tool to evaluate all possible intersections between query words and documents, and computes a relevance weight for each intersection. This relevance weight corresponds to the sum of the weights of words present in the documents.

3 The Sentence Aligner

Parallel text alignment based on cross-language information retrieval consists in building a database of sentences of the target text and considering each sentence of the source text as a "query" to that database (Figure 2).

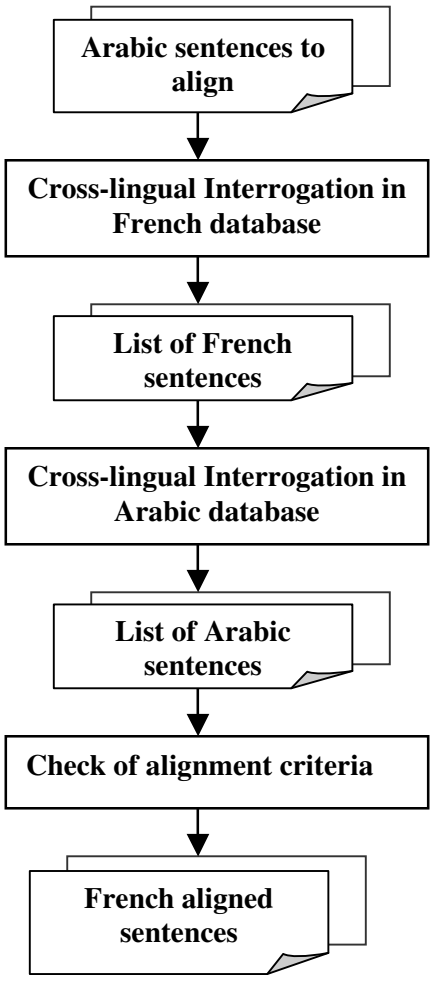

Figure 2. Sentence alignment steps

To evaluate whether the two sentences are translations of each other, we use three criteria:

- Number of common words between the source sentence and the target sentence (semantic intersection) must be higher than 50% of number of words of the target sentence.
- Position of the sentence to align must be in an interval of 10 compared to the position of the last aligned sentence.
- Ratio of lengths of the target sentence and the source sentence (in characters) must be higher or equal than 1.1 (A French character needs 1.1 Arabic characters): *Longer sentences in Arabic tend to be translated into longer sentences in French, and shorter sentences tend to be translated into shorter sentences*.

The alignment process has four steps:

1. Exact match 1-1 alignment: The goal of this step is to obtain an alignment with a maximum precision by using the three criteria: Number of common words between the source sentence and the target sentence; Position of the sentence to align; Ratio of lengths of the target sentence and the source sentence.
2. 1-2 alignment: This alignment consists in merging an unaligned sentence with one preceding or following already aligned sentence. We use to validate this alignment only the first two criteria.
3. 2-1 alignment: The goal of this alignment is to find for the two sentences following an aligned sentence a sentence in the target language taking into account the position of the last aligned sentence. This alignment is validated by using only the first two criteria.
4. Fuzzy match 1-1 alignment: This alignment proposes for the sentence to align the first sentence of the first class returned by the cross-language search engine. This type of alignment is added to take into account alignments which are partially correct (The source sentence is not completely aligned but some of its words are translated).

We describe below the algorithm of the Exact Match 1-1 alignment which is the base of the other aligners. This algorithm uses the functions of the cross-language search engine API.

- PerformCrosslinguageSearch(Query, Corpus, Source language, Target language): returns the set of relevant classes corresponding to the question "Query" in the database "Corpus". Each class is composed of a set of sentences in the target language.
- GetNumberOfCommonWords(Class): returns the number of common words between the source sentence and the target sentence (semantic intersection).
- GetNumberOfWords(Sentence): returns the number of words of a sentence.
- GetNumberOfCharacters(Sentence): returns the number of characters of a sentence.

```
function GetExactMatchOneToOneAlignments(Corpus_Ar, Corpus_Fr)
  for each Arabic sentence P_jAr ∈ Corpus_Ar do
    C_Fr←PerformCrosslinguageSearch(P_jAr, Corpus_Fr, Ar, Fr)
    R←0; Initialize the position of the last aligned sentence.
    for each class C_lFr ∈ C_Fr do
      for each French sentence P_mFr ∈ C_lFr do
        C_Ar←PerformCrosslinguageSearch(P_mFr, Corpus_Ar, Fr, Ar)
        for each class C_qAr ∈ C_Ar do
          for each Arabic sentence P_qAr ∈ C_qAr do
            if P_qAr=P_jAr then
              NM_Fr=GetNumberOfCommonWords(C_lFr);
              NM_Ar=GetNumberOfWords(P_jAr);
              NC_Ar=GetNumberOfCharacters(P_jAr);
              NC_Fr=GetNumberOfCharacters(P_mFr)
              if (NM_Fr ≥ NM_Ar/2) and (R−5 ≤ m ≤ R+5) and (NC_Fr=(1.1)*NC_Ar) then
                The sentence P_mFr is the alignment of the sentence P_jAr;
                R←m
              end if
            end if
          end for
        end for
      end for
    end for
  end for
end function
```

For example, to align the Arabic sentence [4/30] (sentence of position 4 in the Arabic corpus containing 30 sentences) " في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته " (In Italy, the order of things persuaded in an invisible way a majority of electors that time of traditional parties was finished), the exact match 1-1 aligner proceeds as follows:

- The Arabic sentence is considered to be a query to the French sentence database using the cross-language search engine. Retrieved sentences for the two first classes are illustrated in Table 2.

Class	Number of retrieved sentences	Retrieved sentences
1	1	[4/36] En Italie, l'ordre des choses a persuadé de manière invisible une majorité d'électeurs que le temps des partis traditionnels était terminé
2	3	[32/36] Au point que, dès avant ces élections, un hebdomadaire britannique, rappelant les accusations portées par la justice italienne contre M. Berlusconi, estimait qu'un tel dirigeant n'était pas digne de gouverner l'Italie, car il constituait un danger pour la démocratie et une menace pour l'Etat de droit [34/36] Après le pitoyable effondrement des partis traditionnels, la société italienne, si cultivée, assiste assez impassible (seul le monde du cinéma est entré en résistance) à l'actuelle dégradation d'un système politique de plus en plus confus, extravagant, ridicule et dangereux [36/36] Toute la question est de savoir dans quelle mesure ce modèle italien si préoccupant risque de s'étendre demain à d'autres pays d'Europe

Table 2. Retrieved sentences corresponding to the Arabic sentence [4/30]

- Results of cross-language querying show that the sentence [4/36] is a good candidate to alignment. To confirm this alignment, we use the French sentence as a query to the Arabic database. Relevant sentences corresponding to the French query "En Italie, l'ordre des choses a persuadé de

manière invisible une majorité d'électeurs que le temps des partis traditionnels était terminé" are grouped into two classes in Table 3.

Class	Number of retrieved sentences	Retrieved sentences
1	1	[4/30] في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته []
2	3	[26/30] يشكل هؤلاء الرجال اكثر ثلاثية مثيرة للسخرية والتقزز في اوروبا، الى درجة ان احدى المجلات الاسبوعية البريطانية اعتبرت في معرض استعادتها للاتهامات القضائية الموجهة الى السيد برلوسكوني قبل هذه الانتخابات ان مسؤولا من هذا النوع ليس جديرا بحكم ايطاليا وانه يمثل خطرا على الديموقراطية وعلى دولة القانون [] [28/30] وقد تبينت صحة هذه التوقعات المتشائمة، فبعد الانهيار المثير للشفقة للاحزاب التقليدية، شهد المجتمع وف بثقافته ومن دون ان الايطالي المعر يبدي حراكا باستثناء قطاع السينما الذي لجأ الى المقاومة التدهور الراهن لنظام سياسي يعاني المزيد من الغموض والشطط والسخف والخطورة [] [30/30] وكل المسألة تكمن في معرفة الى اي مدى يمكن هذا النموذج الايطالي المثير ان اوروبية للقلق ان ينتشر غداً في بلد اخرى []

Table 3. The two classes corresponding to the French sentence [4/36]

The first proposed sentence is the original one and more of 50% of the words are common to the two sentences. Furthermore, the length ratio between the French sentence and the Arabic sentence is superior than 1.1 and positions of these two sentences in the databases are the same. Therefore, the exact match 1-1 aligner considers the French sentence [4/36] as a translation of the Arabic sentence [4/30].

4 Experimental Results

The sentence aligner has been tested on the MD corpus of the ARCADE II project which is composed of news articles from the French newspaper "Le Monde Diplomatique" (Chiao et al., 2006). This corpus contains 5 Arabic texts (244 sentences) aligned at the sentence level to 5 French texts (283 sentences). The test consisted to build two databases of sentences (Arabic and French) and to consider each Arabic sentence as a "query" to the French database.

To evaluate the sentence aligner, we used the following measures:

$$\text{Precision} = \frac{|A \cap A_r|}{|A|} \text{ and Recall} = \frac{|A \cap A_r|}{|A_r|}$$

A corresponds to the set of alignments provided by the sentence aligner and $\mathbf{A_r}$ corresponds to the set of the correct alignments.

The results we obtained at sentence level (Table 4) show an average precision around 97% and an average recall around 93%. These results do not take into account alignments which are partially correct (Fuzzy match 1-1 alignment).

Parallel Text	Precision	Recall
1	0.969	0,941
2	0,962	0,928
3	0,985	0,957
4	0,983	0,952
5	0,966	0,878

Table 4. Results of alignment at sentence level

Analysis of these results shows that our sentence aligner is not sensitive to missing sentences. This is because the first criterion used by our aligner is not related to surface information (sentence position or sentence length) but on the semantic intersection of these sentences.

Moreover, we have noted that precision depends on the discriminate terms which can occur in the source and target sentences.

5 Conclusion and Perspectives

We have proposed a new approach to sentence alignment based on a cross-language information retrieval model combining different information sources (bilingual lexicon, sentence length and sentence position). The results we obtained show correct precision and recall even when the parallel corpus includes changes in sentence order and missing sentences. This is due to the non-sequential strategy used by the sentence aligner. In future work, we plan to improve the alignment with syntactic structures of source and target sentences and to use the aligned bilingual parallel corpus as a translation memory in a computer-aided translation tool.

References

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

Besançon R., de Chalendar G., Ferret O., Fluhr C., Mesnard O., and Naets H. 2003. Concept-Based Searching and Merging for Multilingual Information Retrieval: In *Proceedings of CLEF-2003*.

Brown P., Lai L., and Mercier L. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of ACL-1991*.

Chiao Y. C., Kraif O., Laurent D., Nguyen T., Semmar N., Stuck F., Véronis J., and Zaghouani W. 2006. Evaluation of multilingual text alignment systems: the ARCADE II project. In *Proceedings of LREC-2006*.

Darwish K. 2002. Building a Shallow Arabic Morphological Analyzer in One Day. In *Proceedings of ACL-2002*.

Debili F. and Zouari L. 1985. Analyse morphologique de l'arabe écrit voyellé ou non fondée sur la construction automatique d'un dictionnaire arabe, *Cognitiva*, Paris.

Debili F. and Sammouda E. 1992. Appariement des Phrases des Textes Bilingues. In *Proceedings of the 14th International Conference on Computational Linguistics*.

Fluhr C., Bisson F., and Elkateb F. 2000. *Parallel text alignment using cross-lingual information retrieval techniques*. Boston: Kluwer Academic Publishers.

Gale W.A. and Church K. W. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of Association for Computational Linguistics*.

Gaussier E. 1995. *Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues*. Ph.D. Thesis, Paris VII University.

Grefenstette G. 1997. *Cross-language information retrieval*. Boston: Kluwer Academic Publishers.

Grefenstette G., Semmar N., and Elkateb-Gara F. 2005. Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications. In *Proceedings of ACL-2005 Workshop*.

Kay M. and Röscheisen M. 1993. *Text-translation alignment*. Computational Linguistics, Special issue on using large corpora, Volume 19, Issue 1.

Larkey L. S., Ballesteros L., and Connel 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*.

Melamed I. D. 1996. A Geometric Approach to Mapping Bitext Correspondence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Papageorgious H., Cranias, L., and Piperidis, S. 1994. Automatic Alignment in Parallel Corpora. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.

Salton G. and McGill M. 1983. *Introduction to Modern Information retrieval*. New York: McGraw Hill.