Using Cross-language Information Retrieval for Sentence Alignment

Nasredine Semmar and 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, christian.fluhr@cea.fr

Cross-language information retrieval consists in providing a query in one language and searching documents in different languages. Retrieved documents are ordered by the probability of being relevant to the user's request with the highest ranked being considered the most 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 to be indexed. This system, designed to work on Arabic, Chinese, English, French, German and Spanish, 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. In the case of Arabic, a clitic stemmer is added to the morphological analyzer to segment the input words into proclitics, simple forms and enclitics. 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 expansion can be in the same language or in different languages. 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. 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 corpora based on the LIC2M 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 aligned bilingual parallel corpora can be used as a translation memory in a computer-aided translation tool.

Keywords: Cross-language information retrieval, linguistic analysis, sentence alignment, bilingual corpora, translation memory.

1. INTRODUCTION

Information retrieval consists in providing relevant documents according to the submitted query, and locating as precisely as possible the most informational parts of these documents. The goal of cross-lingual information retrieval is to find relevant documents that are in a different language from that of the query (Grefenstette 1998). The query terms are translated using bilingual dictionaries.

Sentence alignment consists in mapping sentences of the source language with their translations in the target language. A number of automatic sentence alignment approaches have been proposed (Gale and Church 1991) (Debili and Samouda 1992) (Gaussier 1995) (Planas 1998) (Fluhr 2000).

$$\left[\begin{array}{ll} \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{wa-} \\ \text{COORD\_FORM} & \text{wa-} \\ \text{GLOSS} & \text{and} \\ \text{COORD} & \left\{\begin{array}{l} \left[\begin{array}{ll} \text{PRED} & \text{triyniydAd+u} \\ \text{GLOSS} & \text{Trinidad} \\ \text{DEFINITENESS} & + \\ \text{CASE} & \text{nominative} \end{array}\right] \\ \left[\begin{array}{ll} \text{PRED} & \text{-tuwbAguw} \\ \text{GLOSS} & \text{Tobago} \end{array}\right] \end{array}\right\} \end{array}\right] \end{array}\right]$$

In this paper, we present the LIC2M sentence aligner which is based on cross-language information retrieval techniques. The LIC2M sentence aligner was 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 LIC2M cross-lingual 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. LIC2M CROSS-LINGUAL SEARCH ENGINE

The LIC2M cross-lingual search engine (Besançon and al. 2003) is composed of the following modules (FIGURE 1):

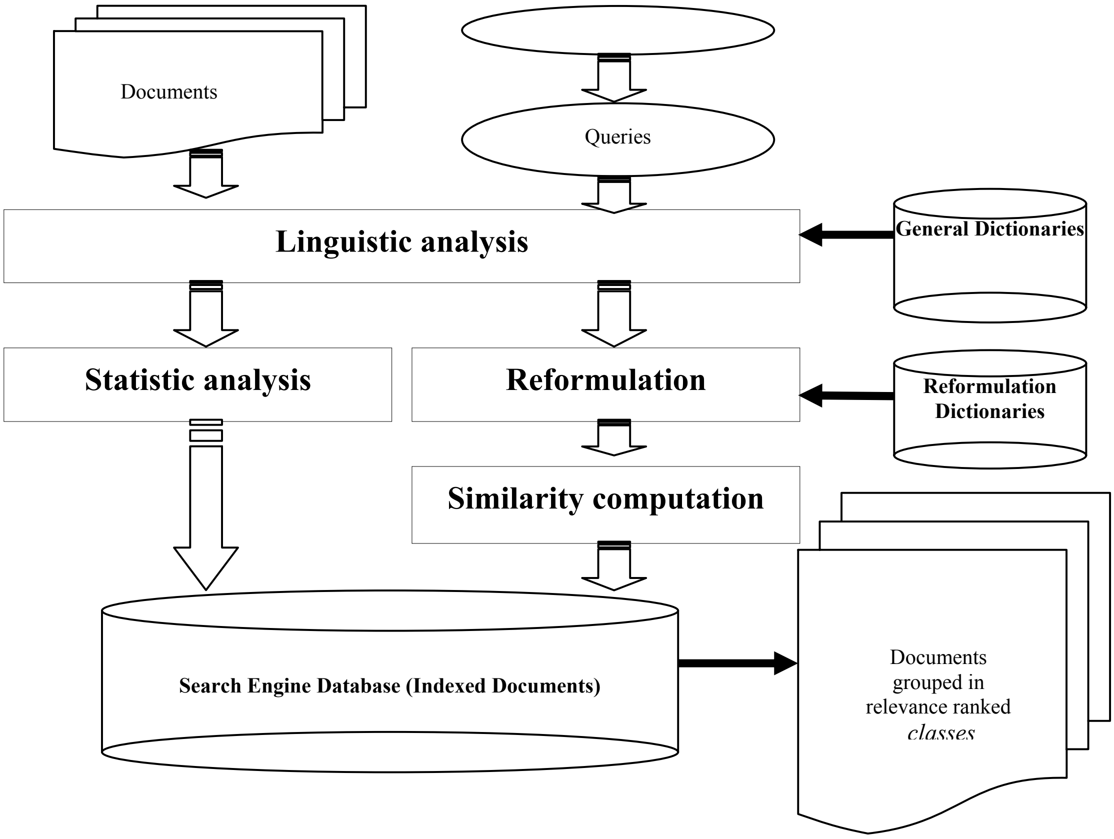
Documents
Queries
Linguistic analysis
General Dictionaries
Statistic analysis
Reformulation
Reformulation Dictionaries
Similarity computation
Search Engine Database (Indexed Documents)
Documents grouped in relevance ranked classes

FIGURE 1: The LIC2M cross-language search engine architecture

- 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, hyponyms, 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 LIC2M linguistic analyzer produces a set of normalized lemmas, a set of named entities and a set of nominal compounds. It is composed of the following modules:

- 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. 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 and Spanish, we added to the morphological analyzer a new processing step: a Clitic stemmer (Larkey and al. 2002) which splits agglutinated words into proclitics, simple forms and enclitics.
- 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.
- 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. 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 (Grefenstette and al. 2005).
- 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.
- A Named Entity recognizer which uses name triggers (e.g., President, lake, corporation, etc.) to identify named entities.

2.2 Statistical Analysis

The statistical analysis consists in attributing 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.

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 can be selected according to the word or the word with a specific part of speech. Currently, the Arabic-French and French-Arabic dictionaries contain around 120 000 entres.

2.3 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 uses a comparison tool to evaluate all possible intersections between query words and documents, and computes a relevance weight for each intersection. It 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 and their weights in order to score the documents.

2.4 User Interface

The LIC2M Search engine is a weighted Boolean system, in which documents are grouped into classes characterized by the same set of concepts as the query. The classes constitute a discrete partition of the database.

TABLE 1 illustrates some classes corresponding to the query “إدارة موارد المياه”. 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	إدارة_موارد_مياه	27 documents: arabic, english, french, spanish
2	موارد_مياه إدارة_موارد	14 documents: arabic, english, french
3	مياه إدارة_موارد	11 documents: arabic, english, spanish
4	إدارة موارد_مياه	24 documents: arabic, english, french, spanish
5	إدارة_موارد	1 documents: english
6	موارد_مياه	9 documents: arabic, english
7	إدارة موارد مياه	37 documents: arabic, english, french, spanish

TABLE 1: First classes corresponding to the query "إدارة موارد المياه"

2. LIC2M 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. To evaluate whether the two sentences are translations of each other, a similarity value is used. This similarity is computed using the semantic overlap and the remainders of the source sentence and the translation. To maximize the precision of alignment, we use three criteria:

- Position of the sentence in the document.
- Number of common words between the source sentence and the target sentence (translation candidate).
- Rate between the length (in terms of number of characters) of the source sentence and the length of the target sentence.

The alignment process has three steps: The first aims at obtaining exact match 1-1 alignments by maximizing the similarity between the source sentence and the translation (use of the three criteria). The second steps tries to find 1-2 or 2-1 alignments by attempting to merge the next unaligned sentence with the previous one already aligned. The goal of the third step is to get the fuzzy match 1-1 alignments, it is achieved by ignoring the sentence order in the parallel corpora and the length rate.

3.1 Exact match 1-1 alignment

The Parallel text is indexed into two databases. These two databases are composed of two sets of ordered sentences, one for each language. The sentence aligner uses a cross-language search to identify the link between the sentence in the source language and the translated sentence in the target language. The exact match 1-1 aligner uses all the criteria.

For example, to align the Arabic sentence **" في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته"** [4/30] (sentence number/number of sentences of the document), 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 LIC2M cross-language search engine.

Retrieved documents for the two first classes are illustrated in TABLE 2.

Class	**Query terms**	**Number of retrieved documents (sentences)**	**Retrieved documents (sentences)**
1	**إيطالِيَا أُد إِقْنَاع طَرِيقَة مَرْئِيّ بَلَغ نِهَايَة طَبِيعَة شَيْء غَالِبِيَّة حِزْب تَقْلِيدِي زَمَن**	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	إيطَالِيَا	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 documents corresponding to the query " في ايطاليا ادت طبيعة الاشياء الى
اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته"

- 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 documents 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	**Query terms**	**Number of retrieved documents**	**Retrieved documents**
1	Italie, persuadé, terminé, temps, parti, traditionnel, ordre, chose, manière, invisible, majorité	1	[4/30] **في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته**

2	Italie	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 الايطالي المعروف بثقافته ومن دون ان 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 للقلق ان ينتشر غداً في بلدان اوروبية اخرى

TABLE 3: The two classes 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"

The first proposed sentence is the original one and more of 50% of the words are common to the two sentences. Furthermore, the length rate between the French sentence and the Arabic sentence is superior than 1.15 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].

3.2 1-2 alignment

To perform 1-2 alignment, we try to merge an unaligned sentence with one preceding or following already aligned sentence. If the intersection increases, we consider the real alignment to be with the concatenated sentences.

For example, the Arabic sentence " **وقد ترسخ هذا الاقتناع في خلاصة مفادها ان النظام السياسي يعاني منذ الثمانينات انحلالا متسارعا**" is aligned with the concatenation of the first French sentence "Cette conviction s'est enracinée dans un constat :" and the second French sentence "le système politique connaît, depuis les années 1980, une dégénérescence accélérée".

3.3 2-1 alignment

We proceed in the same way with each target language unaligned sentence in order to recognize 2-1 alignments.

For example, the concatenation of the first Arabic sentence "**ويؤكد اندرو مونتانيللي:**" and the second Arabic sentence " **يمكن من في عينيه نظر ان يدرك الى اي حد كان مستوى معيشة المسؤولين**

الكبار يتناقض مع تصريحاتهم عن مداخيلهم" is aligned with the French sentence "Quiconque avait des yeux pour voir, a pu affirmer Indro Montanelli, se rendait compte combien le niveau de vie des hauts responsables contrastait avec leurs déclaration de revenus".

3.4 Fuzzy match 1-1 alignment

The fuzzy match 1-1 alignment consists in aligning two sentences which have a low level of similarity. This aligner does not use at all the criteria of precision. It proposes the first document of the first class returned by the LIC2M cross-language search engine.

For example, for the Arabic sentence "الفساد في ايطاليا", the fuzzy match 1-1 aligner proposed the French sentence "La corruption en Italie s'est généralisée et a pris des proportions hallucinantes".

4. EXPERIMENTAL RESULTS

The LIC2M sentence aligner has been tested on the MD corpus of the ARCADE II project which consists of news articles from the French newspaper "Le Monde Diplomatique" (Chiao and al. 2006). The 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 our 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 LIC2M sentence aligner and ***Ar*** corresponds to the set of the correct alignments.

The results at sentence level are summarized by TABLE 4.

Parallel Text	**Number of Arabic sentences**	**Number of French sentences**	**Precision**	**Recall**
1	34	34	0.969	0,941
2	30	36	0,962	0,928
3	80	97	0,985	0,957
4	65	80	0,983	0,952
5	35	36	0,966	0,878

TABLE 4: Results of alignment at sentence level

At sentence level, the precision is around 97% and the recall is around 93%. Moreover, at character level the precision is around 98%. This is due to the fact that the fuzzy match 1-1 aligner copes very well with changes in sentence order and missing sentences.

On the other hand, we realized that precision depends on the discriminate terms which can occur in the source and translated sentences. In fact, a small intersection that contains a discriminate word can confirm the alignment.

5. CONCLUSION AND PERSPECTIVES

We have proposed a new approach to sentence alignment based on cross-language information retrieval technology. 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 results we obtained show that is possible to have with our approach correct precision and recall. In future work, we plan to confirm these results with a large bilingual corpora and improve the alignment by using syntactic structures of source and target sentences.

REFERENCES

Grefenstette, G. (1998) (eds). *Cross-language information retrieval.* 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.

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

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

Planas, E. (1998). *TELA: Structures and Algorithms for Memory-Based Machine Translation.* Ph.D. Thesis, Joseph Fourier University.

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

Besançon R., De Chalendar G., Ferret O., Fluhr C., Mesnard O. and Naets H. (2003). 'The LIC2M's CLEF 2003 system'. In Working Notes for the CLEF 2003 Workshop.

Larkey L. S., Ballesteros L. and Connel (2002) M. E.. '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.

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 ACL-2005 Workshop Proceedings.

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 LREC-2006 Proceedings.