Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons

Nasredine Semmar, Laib Meriama
CEA, LIST, Vision and Content Engineering Laboratory,
18 route du Panorama, Fontenay-aux-Roses, F-92265, France
nasredine.semmar@cea.fr, meriama.laib@cea.fr

Abstract

Translation lexicons are vital in machine translation and cross-language information retrieval. The high cost of lexicon development and maintenance is a major entry barrier for adding new languages pairs. The integration of automatic building of bilingual lexicons has the potential to improve not only cost-efficiency but also accuracy. Word alignment techniques are generally used to build bilingual lexicons. We present in this paper a hybrid approach to align simple and complex words (compound words and idiomatic expressions) from a parallel corpus. This approach combines linguistic and statistical methods in order to improve word alignment results. The linguistic improvements taken into account refer to the use of an existing bilingual lexicon, named entities detection and the use of grammatical tags and syntactic dependency relations between words. The word aligner has been evaluated on the MD corpus of the ARCADE II project which is composed of the same subset of sentences in Arabic and French. Arabic sentences are aligned to their French counterparts. Experimental results show that this approach achieves a significant improvement of the bilingual lexicon with simple and complex words.

1. Introduction

Translation lexicons are a vital component of several Natural Language Processing applications such as machine translation (MT) and cross-language information retrieval (CLIR). The high cost of bilingual lexicon development and maintenance is a major entry barrier for adding new languages pairs for these applications. The integration of automatic building of bilingual lexicons improves not only cost-efficiency but also accuracy. Word alignment approaches are generally used to construct bilingual lexicons (Melamed, 2001).

In this paper, we present a hybrid approach to align simple and complex words (compound words and idiomatic expressions) from parallel text corpora. This approach combines linguistic and statistical methods in order to improve word alignment results.

We present in section 2 the state of the art of aligning words from parallel text corpora. In section 3, the main steps to prepare parallel corpora for word alignment are described; we will focus, in particular, on the linguistic processing of Arabic text. We present in section 4 single and multi-word alignment approaches. We discuss in section 5 results obtained after aligning simple and complex words of a part of the ARCADE II MD (Monde Diplomatique) corpus. Section 6 concludes our study and presents our future work.

2. Related work

There are mainly three approaches for word alignment using parallel corpora:

- Statistical approaches are generally based on IBM models (Brown et al., 1993).
- Linguistic approaches for simple words and compound words alignment use bilingual lexicons and morpho-syntactic analysis on source and target sentences in order to obtain grammatical tags of words and syntactic dependency relations (Debili & Zribi, 1996; Bisson, 2001).
- A combination of the two previous approaches (Daille et al., 1994; Gaussier, 1995; Smadja et al., 1996; Blank, 2000; Barbu, 2004; Ozdowska, 2004). Gaussier (1995) approach is based on a statistical model to establish the French and English word associations. It uses the dependence properties between words and their translations. Ozdowska (2004) approach consists in matching words regards to the whole corpus, using the co-occurrence frequencies in aligned sentences. These words are used to create couples which are starting points for the propagation of matching links by using dependency relations identified by syntactic analysis in the source and target languages.

Machine translation systems based on IBM statistical models do not use any linguistic knowledge. They use parallel corpora to extract translation models and they use target monolingual corpora to learn target language model. The translation model is built by using a word alignment tool applied on a sentence-to-sentence aligned corpus. This model can be represented as a matrix of probabilities that relies target and source words. The Giza++ tool (Och, 2003) implements this kind of approach but its performance is proved only for aligning simple words. Approaches and tools for complex words alignment are at experimental stage (DeNero & Klein, 2008).

3. Pre-processing the bilingual parallel corpus

A bilingual parallel corpus is an association of two texts in two languages, which represent translations of each other. In order to use this corpus in word alignment, two

pre-processing tasks are involved on the two texts: sentence alignment and linguistic analysis.

3.1 Sentence alignment

Sentence alignment consists in mapping sentences of the source language with their translations in the target language. A number of sentence alignment approaches have been proposed (Brown et al., 1991; Gale & Church, 1991; Kay & Röscheisen, 1993).

Our approach to align the sentences of the bilingual parallel corpus combines different information sources (bilingual lexicon, sentence length and sentence position) and is based on cross-language information retrieval which 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 (Semmar & Fluhr, 2007). This approach uses a similarity value to evaluate whether the two sentences are translations of each other. This similarity is computed by the comparator of the cross-language search engine and consists in identifying common words between source and target sentences. This search engine is composed of a deep linguistic analysis, a statistical analysis to attribute a weight to each word of the sentence, a comparator and a reformulator to translate the words of the source sentence in the target language by using a bilingual lexicon.

In order to refine the result of alignment, we used the following three criteria:

- Number of common words between the source sentence and the target sentence (semantic similarity) 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 three steps:

- Exact match 1-1 alignment: In this step, the similarity between the source sentence and the target sentence is maximized by using the three criteria mentioned above.
- 1-2 or 2-1 alignments: The goal of this step is to attempt to merge the next unaligned sentence with the previous one already aligned. To confirm 1-2 or 2-1 alignments, we use only the first two criteria.
- Fuzzy match 1-1 alignment: This step consists in aligning two sentences with a low level of similarity. This aligner does not use the three criteria.

The parallel corpus 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 (Figure 1).

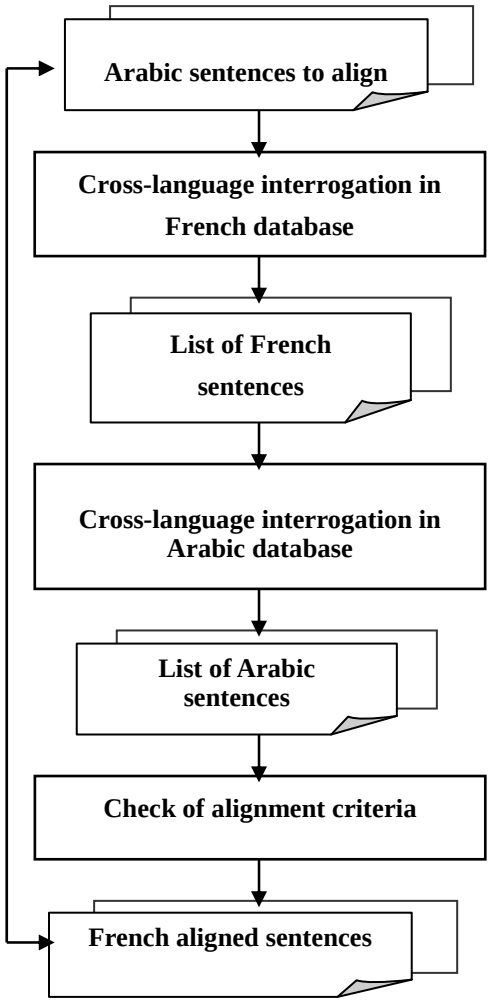

Figure 1: Sentence alignment steps.

3.2 Linguistic analysis

The linguistic analysis produces a set of normalized lemmas, a set of named entities and a set of compound words with their grammatical tags. This analysis is built using a traditional architecture involving separate processing modules:

- A morphological analyzer which looks up each word in a general full form dictionary. If these words are found, they are associated with their lemmas and all their grammatical tags. For Arabic agglutinated words which are not in the full form dictionary, a clitic stemmer was added to the morphological analyzer. The role of this stemmer is to split 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. These rules can recognize contiguous expressions as "البَيْت الأبْيَض" (the white house).

- 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 sequences are generated from a manually annotated training corpus. They are extracted from a hand-tagged corpora. 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 sequence. If no bigrams allow completing the path, the word is left undisambiguated. The following example shows the result of the linguistic analysis after Part-Of-Speech tagging of the Arabic sentence "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 has persuaded in an invisible manner a majority of voters that the time of traditional political parties was completed). Each word is represented as a string (token) with its lemma and morpho-syntactic tag (token: lemma, morpho-syntactic tag).

(1) في: فِي, Preposition
(2) ايطاليا: إِيطَالِيَا, Proper Noun
(3) ادت: أدَّ, Verb
(4) طبيعة: طَبِيعَة, Common Noun
(5) ال: ال, Definite Article
(6) اشياء: شَيْء, Common Noun
(7) الى: إِلَى, Preposition
(8) اقناع: إِقْنَاع, Common Noun
(9) غالبية: غَالِبِيَّة, Common Noun
(10) ال: ال, Definite Article
(11) ناخبين: نَاخِب, Common Noun
(12) في: فِي, Preposition
(13) طريقة: طَرِيقَة, Common Noun
(14) غير: غَيْر, Adverb
(15) مرئية: مَرْئِيّ, Adjective
(16) ب: بِ, Preposition
(17) أن: أنَ, Conjunction
(18) زمن: زَمَن, Common Noun
(19) ال: ال, Definite Article
(20) احزاب: حِزْب, Common Noun
(21) ال: ال, Definite Article
(22) تقليدية: تَقْلِيدِيّ, Adjective
(23) قد: قَد, Preposition
(24) بلغ: بَلَغَ, Verb
(25) نهايت: نَهَايَة, Common Noun
(26) ه: ه, Pronoun

- A syntactic analyzer which is used to split graph of words into nominal and verbal chains and recognize dependency relations 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 a 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 "نقل المياه" (water transportation), the syntactic analyzer considers this nominal chain as a compound word "نقل_مياه" composed of the words "نقل" (transportation) and "مياه" (water).
- A named entity recognizer which uses name triggers to identify named entities. For example, the expression "الأوَّل مِن شَهْر مَارس" (The first of March) is recognized as a date and the expression "قطر" (Qatar) is recognized as a location.
- A module to eliminate empty words which consists in identifying words that should not be used as search criteria and removing them. These empty words are identified using only their Part-Of-Speech tags (such as prepositions, articles, punctuations and some adverbs). For example, the preposition "ل" (for) in the agglutinated word "لنقل" (for transportation) is considered as an empty word.
- A module to normalize words by their lemmas. In the case the word has several lemmas, only one of these lemmas is taken as normalization. Each normalized word is associated with its morpho-syntactic tag. For example, normalization of the word "أنابيب" (pipelines) which is the plural of the word "أنْبُوب" (pipeline) is represented by the couple (أنْبُوب, Noun).

4. Word alignment

Our approach to align simple and complex words adapts and enriches the methods developed by:

- (Debili & Zribi, 1996) (Bisson, 2001) which consist to use, in one hand, a bilingual lexicon and the linguistic properties of named entities and cognates to align simple words, and on the other hand, syntactic dependency relations to align complex words.
- (Giguet & Apidianaki, 2005) which consist to use sequences of words repeated in the bilingual corpora and their occurrences to align compound words and idiomatic expressions.

4.1 Single-word alignment

Single-word alignment is composed of the following steps:

- Alignment using the existing bilingual lexicon.
- Alignment using the detection of named entities.
- Alignment using grammatical tags of words.
- Alignment using Giza++.

4.1.1. Bilingual lexicon look-up

Alignment using the existing bilingual lexicon consists in extracting for each word of the source sentence the appropriate translation in the bilingual lexicon. The result of this step is a list of lemmas of source words for which one or more translations were found in the bilingual lexicon. The Arabic to French lexicon used in this step contains 124 581 entries.

Table 1 shows results of this step for the Arabic sentence " في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته" and its French translation "En Italie, l'ordre des choses a persuadé de manière invisible une majorité d'électeurs que le temps des partis traditionnels était terminé".

Lemmas of words of the source sentence	Translations found in the bilingual lexicon
شَيْء	chose
غَالِبِيَّة	majorité
نَاخِب	électeur
يقَةطَر	manière
زَمَن	temps
حِزْب	parti
تَقْلِيدِيّ	traditionnel

Table 1: Single-word alignment with the existing bilingual lexicon.

4.1.2. Named entities detection

For those words that are not found in the bilingual lexicon, the single-word aligner searches named entities present in the source and target sentences. For example, for the previous Arabic sentence and its French translation, the single-word aligner detects that the Arabic word "إيطالِيَا" (Italy) and the French word "Italie" are named entities of the type "Location". However, this first step can produce alignment errors in the case the source and target sentences contain several named entities. To avoid these errors, we added a criterion related to the position of the named entity in the sentence.

4.1.3. Grammatical tags matching

If for a given word no translation is found in the bilingual lexicon and no named entities are present in the source and target sentences, the single-word aligner tries to use grammatical tags of source and target words. This is especially the case when the word to align is surrounded with some words already aligned. For example, because the grammatical tags of the words "طبيعَة" and "ordre" are the same (Noun) and "طبيعَة" is surrounded with the words "إيطالِيَ" and "شَيْء" which are already aligned in the two previous steps, the single-word aligner considers that the lemma "ordre" is the translation of the lemma "طبيعَة".

4.1.4. Giza++ alignment

For those words that are not found in the bilingual lexicon and are not aligned by named entities detection or grammatical tags matching, the single-word aligner uses results obtained with the Giza++ aligner from the bilingual parallel corpus. For example, Giza++ founds that the French word "persuasion" is a translation of the Arabic word "اقناع" despite the fact that this word does not belong to the French sentence "En Italie, l'ordre des choses a persuadé de manière invisible une majorité d'électeurs que le temps des partis traditionnels était terminé". In addition, this word has not vowels because it is taken directly from the parallel corpus. Table 2 illustrates results after running the four steps of single-word alignment.

Lemmas of words of the source sentence	Translations returned by single-word alignment
إيطالِيَا	Italie
طبيعَة	ordre
شَيْء	chose
اقناع	persuasion
غَالِبِيَّة	majorité
نَاخِب	électeur
طريقَة	manière
زَمَن	temps
حِزْب	parti
تَقْلِيدِيّ	traditionnel

Table 2: Single-word alignment results.

4.2 Multi-word alignment

The results obtained by the current tools for aligning words from parallel corpora are limited either to the extraction of bilingual simple words from specialized texts or to the extraction of bilingual noun phrases from texts related to the general field. These limitations are due to the fact that the extraction of compound words is more difficult than the extraction of simple words. The following examples illustrate some difficulties encountered when aligning compound words:

- A compound word is not automatically translated with a compound word. For example, the Arabic compound word "إعلام آلِيٌّ" is translated as a single word in French "informatique".
- The translation of a compound word is not always obtained by translating its components separately. For example, the French translation of the Arabic compound word "تَحْتَ التَّسْدِيدِ" is not "sous le règlement" but "en cours de règlement".
- A same compound word can have different forms due to the morphological, syntactic and semantic changes. These changes must be taken into account in the alignment process. For example, the Arabic compound words "إدارة موارد المياه" and " إدارة الموارد المائية" have the same French translation "gestion des resources en eau".

Our multi-word alignment approach is composed of the following steps:

- Alignment of compound words that are translated literally from one to the other.
- Alignment of idiomatic expressions and compound words that are not translated word for word.

4.2.1. Compound words alignment

Compound words alignment consists in establishing correspondences between the compound words of the

source sentence and the compound words of the target sentences. First, a syntactic analysis is applied on the source and target sentences in order to extract dependency relations between words and to recognize compound words structures. Then, reformulation rules are applied on these structures to establish correspondences between the compound words of the source sentence and the compound words of the target sentence. For example, the rule *Translation(*A.B*) = Translation(*A*).Translation(*B*)* allows to align the Arabic compound word "حزْب تَقْلِيدِيّ" with the French compound word "parti traditionnel" as follows:

Translation(حزْب.تَقْلِيدِيّ) =
Translation(حزْب).Translation(تَقْلِيدِيّ) = parti. traditionnel

In the same manner, this step aligns the compound word "طَبِيعَة_شَيْء" with the compound word "ordre_chose" even if the word "ordre" is not proposed as a translation of the word "طَبِيعَة" in the bilingual lexicon.

4.2.1. Idiomatic expressions alignment

In order to translate missed compound words and idiomatic expressions, we used a statistical approach which consists in:

- identifying the sequences of words which are candidate for the alignment: for the two texts of the bilingual corpus, we compute the sequences of repeated words and their number of occurrences.
- representing these sequences with vectors: for each sequence, we indicate numbers of segments in which the sequence appears.
- aligning the sequences: for each sequence of the source text and each sequence of the target text, we estimate the value of the translation relation with the following formula:

$$\cos(x_i, y_i) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2} \times \sqrt{\sum y_i^2}}$$

This step results in a list of single words, compound words and idiomatic expressions of the source sentence and their translations. For example, for the previous Arabic sentence and its French translation, the multi-word aligner founds that the expression "manière invisible" is a translation of the Arabic expression "طريقة غير مرئية".

4.3 Cleaning the bilingual lexicon

The various approaches described in this paper to align simple and complex words use different tools for terminology extraction and dependency syntactic analysis. Each of these tools can be a source of noise because of errors that can be produced by the modules that compose them (POS tagging, lemmatization …). Therefore, these approaches inevitably produce incorrect matches between the words of source text and the words of target text. It thus becomes important to remove incorrect entries and retain only the correct words in the bilingual lexicons built or updated automatically by these methods.

We have established a score for each type of alignment to facilitate the cleaning process of the bilingual lexicon built or updated automatically from the parallel corpus:

- A link alignment between single words found in the bilingual corpus and validated in the bilingual dictionary has a score equal to 1.
- A link alignment between single words found by the detection of named entities (proper nouns and numerical expressions) has a score equal to 0.99.
- A link alignment between single words found by matching grammatical tags has a score equal to 0.98.
- A link alignment between single words produced by GIZA++ has a score equal to 0.97.
- A link alignment between compound words that are translated literally from one to the other has a score equal to 0.96.
- A link alignment between compound words that are not translated word for word or idiomatic expressions has a score equal to 0.95.

Table 3 presents results after running all the steps of word alignment process for simple and complex words.

Simple and complex words of the source sentence	Translations returned by word alignment	Score
إيطَالِيَا	Italie	0.99
طَبِيعَة	ordre	0.98
شَيْء	chose	1
اقناع	persuasion	0.97
غَالِبِيَّة	majorité	1
نَاخِب	électeur	1
طَرِيقَة	manière	1
زَمَن	temps	1
حِزْب	parti	1
تَقْلِيدِيّ	traditionnel	1
غَالِبِيَّة_نَاخِب	majorité_électeur	0.96
حِزْب_يّتَقْلِيد	parti_traditionnel	0.96
زَمَن_حِزْب_تَقْلِيدِيّ	temps_parti_traditionnel	0.96
طَبِيعَة_شَيْء	ordre_chose	0.96
طريقة غير مرئية	manière invisible	0.95

Table 3: Single-word and multi-word alignment results.

5. Experimental results

The word 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" (Veronis et al., 2008). The corpus contains 5 Arabic texts (244 sentences) aligned at the sentence level to 5 French texts (283 sentences). The performance of the word aligner is presented in Table 4.

Precision	Recall	F-measure
0.85	0.80	0.82

Table 4: Word alignment performance.

Analysis of the alignment results of the previous sentence (Table 3) shows, in one hand, that 10 simple words (among 14), 4 compound words and 1 idiomatic expression are correctly aligned, and on other hand, 7 simple words are aligned with the bilingual lexicon, 1 simple word is aligned with named entities detection, 1 simple word is aligned by using grammatical tag matching and 1 simple word is aligned with Giza++.

For the whole corpus, 53% of words are aligned with the bilingual lexicon, 9% are aligned with named entities detection, 15% are aligned by using grammatical tags and 4% are aligned as compound words or idiomatic expressions. Consequently, 28% of the words of the source sentence and their translations are added to the bilingual lexicon.

6. Conclusion

In this paper, we have presented a hybrid approach to word alignment combining statistical and linguistic sources of information (bilingual lexicon, named entities detection, use of grammatical tags and syntactic dependency relations, number of occurrences of word sequences). The results we obtained showed that this approach improves word alignment precision and recall, and achieves a significant enrichment of the bilingual lexicon with simple and complex words. In future work, we plan to develop strategies and techniques, in one hand, to filter word alignment results in order to clean the bilingual lexicons built or updated automatically, and on other hand, to improve the recall of the statistical approach by using the existing bilingual lexicon and the results of the morpho-syntactic analysis of the parallel corpus.

7. Acknowledgements

This research work is supported by WEBCROSSLING (ANR - Programme Technologies Logicielles - 2007) and MEDAR (Support Action FP7 – ICT – 2007 - 1) projects.

8. References

Barbu, A.M. (2004). Simple linguistic methods for improving a word alignment. In *Proceedings of the 7th International Conference on the Statistical Analysis of Textual.*

Bisson, F. (2000). U Méthodes et outils pour l'appariement de textes bilingues. Thèse de Doctorat en Informatique. Université Paris VII.

Blank, I. (2000). *Parallel Text Processing : Terminology extraction from parallel technical texts.* Dordrecht: Kluwer.

Brown, P.F., Mercier, L. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of ACL 1991*.

Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R. L. (1993). The mathematics of statistical machine translation : parameter estimation. *Computational Linguistics* 19(3).

Daille, B., Gaussier, E., Lange, J. M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94).*

Debili, F., Zribi, A. (1996). Les dépendances syntaxiques au service de l'appariement des mots. In *Proceedings of the 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle.*

DeNero, J., Klein, D. (2008). The Complexity of Phrase Alignment Problems. In *Proceedings of the of ACL 2008.*

Gale, W.A., 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., Lange, J.M. (1995). Modèles statistiques pour l'extraction de lexiques bilingues. *Traitement Automatique de la Langue* 36.

Giguet, E., Apidianaki, M. (2005). Alignement d'unités textuelles de taille variable. In *Proceedings of the 4èmes Journées de la Linguistique de Corpus.*

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

Melamed, I.D. (2001). *Empirical Methods for Exploiting Parallel Texts.* MIT Press.

Och, F.J. (2003). *GIZA++: Training of statistical translation models.* MIT Press http://www.fjoch.com/GIZA++.htm.

Ozdowska, S. (2004). Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. In *Proceedings of the 11ème conférence TALN-RECITAL.*

Smadja, F., Mckeown, K., Hatzivassiloglou, V. (1996). Translation Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics* 22(1).

Semmar, N., Fluhr, C. (2007). Arabic to French Sentence Alignment: Exploration of A Cross-language Information Retrieval Approach. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources.*

Veronis, J., Hamon, O., Ayache, C., Belmouhoub, R., Kraif, O., Laurent, D., Nguyen, T. M. H., Semmar, N., Stuck, F., Zaghouani, W. (2008). *Arcade II Action de recherche concertée sur l'alignement de documents et son évaluation.* Chapitre 2, Editions Hermès.