International Journal of Scientific & Engineering Research Volume 9, Issue 9, September-2018 1961
ISSN 2229-5518

Alignment Methods For Parallel Texts

Naoual BOUHRIM, Lahbib ZENKOUAR

Abstract— This paper aims to present a review of the methods and techniques used to align parallel texts. The alignment methods included in this paper are exposed to three categories; sentences alignments, words Alignment, and Phrase-based alignment. The main purpose is to establish a benchmarking between the different methods using the phrase-based alignment, the result of this work will help us in the next researches to choose the right approach for the Automatic Language Translation (French / Amazighe)..

Index Terms— ALT, Corpus, parallel text, Sentence alignment, text alignment.

1 Introduction

TO improve the communication between different parties, each with its own language, many types of research in the field of Automatic Language Translation try to find best solutions that will help us to have a high quality of translation with optimizing resources and time.

The text alignment between two or more languages is a preliminary step for the automatic translation.

For this, our job is to identify the different methods used in the state of the art, for the Automatic Language Translation.

In this paper, the first section reserved to the terminology, the second will be devoted to the methods of phrase-based alignment and finally, the third section will be dedicated to the alignment application.

2 Terminology

2.1 Parallel text

In the ALT area, the parallel texts are represented as a set of his original texts and their translations.

According to Véronis, 2000 [1], these parallel texts represent a source of ambiguity depending on the domain where this terminology is used.

Furthermore, it's found in the literature the terminology of "comparable Texts" which represents sets of text, Provided that none is a translation of one of them.

2.2 Proposal

It is an entity smaller than the "Sentence" and bigger than a "Word", generally surrounded by graphic separators.

2.3 Sentence

According to Simard (1998) [2], the sentence is a sequence of syntactically autonomous words, which ends with «.».

- *Naoual BOUHRIM, Phd Student, LEC Departement, MOHAMMADIA SCHOOL OF ENGINNERS MOHAMMED V UNIVERSITY OF RABAT, MOROCCO (naoualensias@gmail.com)*
- *Prof Lahbib ZENKOUAR, LEC Departement, MOHAMMADIA SCHOOL OF ENGINNERS MOHAMMED V UNIVERSITY OF RABAT, MOROCCO (Lahbib.zenkouar@gmail.com)*

2.4 Alignement

In this work, the alignment is defined as being at the same time the operation of alignment and its result.

2.5 Alignement operation

It is about a process which receives several texts $T^{l_1}, \dots, T^{l_n}$ drafted in various languages $l_1, \dots, l_n$ having for result, a the list of sets $L = \{P_1, \dots, P_m\}$ Established each of the element $\left(E_i^{l_1}, \dots, E_i^{l_n}\right)$

2.6 The element to be aligned

The mathematical formulation of the element to be aligned can define itself in the following way:

$$E_i^l = \{u_j^l | u_j^l \in T^l \quad 1 \le j \le k\} \qquad avec\ 1 \le i \le n$$

Let F^l the assembly of these elements E_i^l, we will find :

$$F^l = \bigcup_{i=1}^{n} E_i^l = T^l$$

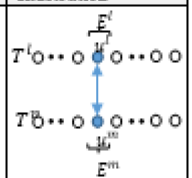
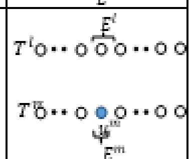
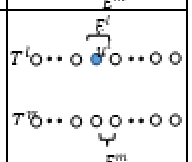
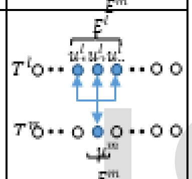
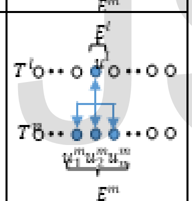
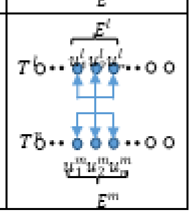
2.7 Pearl

It is a terminology proposed by Brown and al. (1991) [3], which defines the element resulting from the alignment of two texts of entry T^l, T^m

$$P_i^{lm} = \left(E_p^l, E_q^m\right) E_p^l \in F^l, \qquad E_q^m \in F^l$$

N.B : E_p^l *and* E_q^m can have null values according to Gale and Church (1993) [4], There are six models of pearl in the literature:

TABLE 1
CLASSIFICATION OF PEARLS

Model	Description	Illustration
Substitution *Pearl-lm*	Result of a substitution of a unity u^l by a unity u^m $P = (\{u^l\}, \{u^m\})$	
Abolition *Pearl-l*	Result of an abolition of a unity u^l $P = (\{u^l\}, \{\})$	
Insertion *Pearl-m*	Result of an insertion of a unity u^m $P = (\{\}, \{u^m\})$	
Contraction *Pearl-l⁺m*	Result of a contraction of a unity u^m more than one unity $u_1^l, u_2^l, \ldots, u_n^l$ $P = (\{u_1^l, u_2^l, \ldots, u_n^l\}, \{u^m\})$	
Expansion *Pearl-lm⁺*	Result of an extension of a unity u^l more than one unity $u_1^m, u_2^m, \ldots, u_n^m$ $P = (\{u^l\}, \{u_1^m, u_2^m, \ldots, u_n^m\})$	
Fusion *Pearl-l⁺m⁺*	Result of a fusion of several units $u_1^l, u_2^l, \ldots, u_n^l$ more than a unity $u_1^m, u_2^m, \ldots, u_n^m$ $P = (\{u_1^l, u_2^l, \ldots, u_n^l\}, \{u_1^m, u_2^m, \ldots, u_n^m\})$	

2.8 Alignement Result

An alignment L, result of the alignment operation's, consists of a number t of pearls P :

$$L^{lm} = \{P_i^{lm} | P_i^{lm} \in (F^l, F^m) \qquad 1 \leq i \leq t\}$$

3 METHODS OF PHRASE ALIGNEMENT

The first work, which founded in the literature, belongs to Kay and Röscheisen (1988) [5].The majority of the works, which succeed the latter, based generally on three
Hypothesis:

- The order of the sentences is identical.
- Generally, the alignment is of the order of 1: 1.

He is rare to make an abolition or an insertion of a unity.
The following part, present the detail the approaches used for every kind of alignment:

3.1 A method based on the information of lexical correspondence

Kay and Kay and Röscheisen (1993) [6] proposed the first method based on the lexical correspondence.

The procedure of alignment consists of four operations, namely:

- The construction of the table« Word-Sentence Index » (WSI).
- The construction of the table « Alignable Sentence Table » (AST).
- The construction of the table « Word Alignment Table » (WAT).
- The construction of the table « Sentence Alignment Table » (SAT).

The majority of the works using this method are based on two hypothesis:

Hypothesis 1: *Correspondence of contents*

Let's take two parallel texts T^A and T^Bsuch as :

* T^A is constituted by the sentences:

$$P_1^A = u_1^A u_2^A u_3^A \,;\, P_2^A = u_4^A u_5^A \,;\, P_1^A = u_6^A u_7^A u_8^A \,;$$

* T^B is constituted by the sentences:

$$P_1^B = u_1^B u_2^B u_3^B \,;\, P_2^B = u_4^B u_5^B \,;\, P_1^B = u_6^B u_7^B u_8^B \,;$$

If the various elements have the correspondences below:

$$u_1^A \leftrightarrow u_1^B \,;\, u_2^A \leftrightarrow u_2^B \,;\, u_3^A \leftrightarrow u_3^B$$

$$u_4^A \leftrightarrow u_4^B \,;\, u_5^A \leftrightarrow u_5^B$$

$$u_6^A \leftrightarrow u_6^B \,;\, u_7^A \leftrightarrow u_8^B \,;\, u_9^A \leftrightarrow u_9^B$$

The sentences are considered below are aligned:

$$P_1^A \leftrightarrow P_1^B$$

$$P_2^A \leftrightarrow P_2^B$$

$$P_3^A \leftrightarrow P_3^B$$

Hypothesis 2: *Diagonalisation of the alignment*

This hypothesis consists in creating the correspondence of the sentences according to the diagonal of both texts T^A and T^B:

$$\begin{matrix} P_1^A \\ P_2^A \\ P_3^A \end{matrix} \begin{pmatrix} \cdot & \cdot & \blacksquare \\ \cdot & \blacksquare & \cdot \\ \blacksquare & \cdot & \cdot \end{pmatrix}$$

$$P_1^B \quad P_2^B \quad P_3^B$$

$$Diagonal = \{(P_1^A, P_1^B), (P_2^A, P_2^B), (P_3^A, P_3^B)\}$$

ISSN 2229-5518

3.1.1 Table « Word-Sentence Index » (WSI)

The objective of this table is to define the cases of an element E which not have the same form.

We define lemmas as being sub-chains with prefixes or suffixes. To divide these sub-chains, a called method sorts out proposed by Knuth, 1997 [7] allows recognizing initial or final sequences common to several words.

Let us consider the following list:

$$\left.\begin{array}{ll} Abcdr & => Abc + dr \\ Abckfg & => Abc + kfg \\ Abckl & => Abc + kl \\ abcm & => Abc + m \end{array}\right\} \textbf{\textit{ABC is our lemma}}$$

Example :
The lemma of the cases: *Nouvelles*, *nouvelle*, *nouveaux*, *nouveau* est ***nouve.***

3.1.2 Table « Alignable Sentence Table » (AST)

As previously indicated, the alignment of the sentences of two parallel texts can be deducted thanks to the calculation of the diagonal; the obtained result is a shape of correspondence between a sentence of the source text with a sentence of the target text.

In the reality, a sentence can be aligned with several sentences, and that needs more calculation to define these alignments.

3.1.3 Table « Word Alignment Table » (WAT)

This table contains the pairs of elements having a raised rate of similarity of distribution. The case of these following two texts has been taken for illustration:

T^A	T^B
$abc - PhraseP_1^A$	$\alpha\beta\gamma - PhraseP_1^B$
$ade - PhraseP_2^A$	$\alpha\delta\theta - PhraseP_2^B$
$cbf - PhraseP_3^A$	$\gamma\beta\varphi - PhraseP_3^B$
$abf - PhraseP_4^A$	$\alpha\beta\varphi - PhraseP_4^B$

The diagonal of $T^A \times T^B$ is :

$$Diagonal = \{(P_1^A, P_1^B), (P_2^A, P_2^B), (P_3^A, P_3^B), (P_4^A, P_4^B)\}$$

The distribution of every element of the example is:

$Distribution\ de\ a = \{P_1^A, P_2^A, P_4^A\}$
$Distribution\ de\ b = \{P_1^A, P_3^A, P_4^A\}$
$Distribution\ de\ c = \{P_1^A, P_2^A\}$
$Distribution\ de\ d = \{P_2^A\}$
$Distribution\ de\ e = \{P_2^A\}$
$Distribution\ de\ f = \{P_3^A, P_4^A\}$
$Distribution\ de\ \alpha = \{P_1^B, P_2^B, P_4^B\}$
$Distribution\ de\ \beta = \{P_1^B, P_3^B, P_4^B\}$
$Distribution\ de\ \gamma = \{P_1^B, P_2^B\}$
$Distribution\ de\ \delta = \{P_2^B\}$
$Distribution\ de\ \theta = \{P_2^B\}$
$Distribution\ de\ \varphi = \{P_3^B, P_4^B\}$

Among the pairs susceptible to be aligned, b and ß are founded to confirm the correspondence we have to calculate their similarity, to be made this we must:

1- Calculate the Cartesian product of both distributions:
$R = \{(P_1^A, P_1^B), (P_1^A, P_3^B), (P_1^A, P_4^B), (P_3^A, P_1^B), (P_3^A, P_3^B), (P_3^A, P_4^B), (P_4^A, P_1^B), (P_4^A, P_3^B), (P_4^A, P_4^B)\}$

2- Calculate the result of the intersection between the Cartesian product and the diagonal:

$$R \cap Diagonale = \{(P_1^A, P_1^B), (P_3^A, P_3^B), (P_4^A, P_4^B)\}$$

3- Calculate the similarity S according to the coefficient of Dice (van Rijsbergen, 1979 [8]).

$$S = \frac{2|R \cap Diagonale|}{|b\ Distribution| + |\beta\ Distribution|} = 1$$

Because the value of “1” is the maximal value which the similarity can reach, it can be concluded that elements b and ß are corresponding elements.

There are other methods which exist in the state of the art and which define the similarity S based on the frequency of the words, Example:

- Method of Kay ;
- Method of Gale (Gale& Church (1991) [9];
- Method of BACCS (Fung & Church, 1994) [10].

3.1.4 Table « Sentence Alignment Table » (SAT)

This table contains the sentences that can be aligned with a corresponding number of elements.

TABLE 2
TABLE SAT

Pair of sentence	Corresponding elements	
	Group of elements	*Number*
(P_1^A, P_1^B)	$(a,\alpha),(b,\beta),(c,\gamma)$	3
(P_2^A, P_2^B)	$(a,\alpha),(d,\delta),(\varepsilon,\theta)$	3
(P_3^A, P_3^B)	$(b,\beta),(c,\gamma),(f,\varphi)$	3
(P_4^A, P_4^B)	$(a,\alpha),(b,\beta),(f,\varphi)$	3

As the number of correspondence is big, as the alignment of two sentences is justified.
In our case, we can confirm the alignment between the sentences $P_1^A et P_1^B$- $P_2^A et P_2^B$- $P_3^A et P_3^B$- $P_4^A et P_4^B$.

3.1.5 Advantages & Disadvantages

The main advantage of these types of methods lives in the independence of the languages to be handled. but the major inconvenience which we can meet, is the problem of matching of the words having small frequencies (Example: If we have 2 words which have a frequency equal to 1 in the source text and the single word in the target text with the same frequency (1), in this case the matching is impossible).

3.2 Alignement Method based on length correlation.

The approach of these types of methods based on the hypothesis saying that both parallel texts have the same number of sentences or a logical report between the lengths of both parallel texts.
The approach of these types of methods based on the hypothesis saying that both parallel texts have the same number of sentences or a logical report between the lengths of both parallel texts.
Among the most known methods, we quote the method basing itself on the length of the sentences proposed in the works of [Brown on 1991] [3] and [Mange 1993] [4]. This method considers that the length of the sentences in the source text and that of their translations in the target language are correlated and the second method is based on the length of segments presented in the searches (researches) of [VERONIS, LANGLAIS on 2000].

3.2.1 Method of Brown

This method considers a text as being a sequence of lengths of sentences (number of the graphic words) more the markers (scorers) of paragraph (Carriage return - ¶).

According to Brown, we can represent a text in the French language in the following way: $T_f = 18_f 22_f 4_f ¶_f$

This representation illustrates a case of a paragraph written in French language compound of three sentences established respectively by 18, 22 and 4 graphic words.

This is an example of an alignment between two texts (French and amazighe):

$T_f = 18_f 22_f 4_f ¶_f$
$T_a = 16_a 24_a 3_a ¶_a$

There are eight possibilities of pearls in the case of the method of Brown, namely:

- *Pearl-fa* : A French sentence and an amazighe sentence ;
- *Pearl-f*: A French sentence and no sentence on amazighe;
- *Pearl-a*: No French sentence and a sentence on amazighe;
- *Pearl-ffa*: Tow French sentence and one amazighe ;
- *Pearl-faa*: A French sentence and two sentences amazighe;
- *Pearl-¶f*: A marker of French paragraph and no marker of amazighe ;
- *Pearl-¶a*: No marker of French paragraph and a marker of amazighe;
- *Pearl-¶f¶a* : A marker of French paragraph and a marker of amazighe;

If the correct alignment corresponds to:
{(1st French sentence -18_f,1st amazighe sentence -16_a),
(2nd and 3rd French sentence -$22_f 4_f$,2nd amazighe sentence -24_a),
(3rd amazighe sentence -3_a),
(Marker of the paragagraph -$¶_f$, Marker of the paragraph -$¶_a$)}.

He is represented as follows:

$$Pearl - fa, Pearl - ffa, Pearl - a, Pearl - ¶_f ¶_a$$

The method of Brown is based on a model HMM (*Hidden Markov Model*) which distorts between two random processes Generation of pearls and Calculation of the probability of pearls.

3.2.2 Method of Gale

Contrary to the method of Brown, the method of Mange calculates the length of a sentence according to the number of characters and not of the number of words.

Gale added the hypothesis, where we can have a combination established by two sentences of every text, to the hypothesis proposed by Brown. For that purpose, Gale defined six models:

-Substitution (1-1) ;
-Abolition (1-0) ;
-Insertion (0-1) ;
-Contraction (2-1) ;
-Expansion (1-2) ;
-Fusion (2-2).

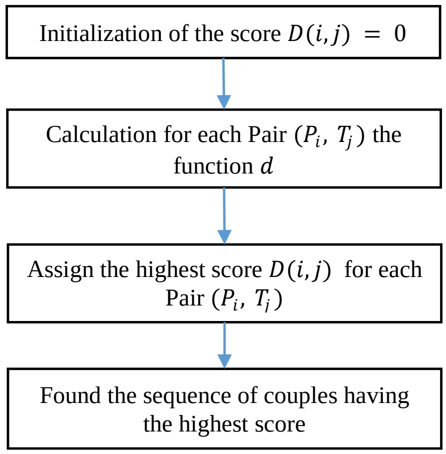


Fig. 1. The algorithm of Gale

With:
$\boldsymbol{P_i}$**:** The i^{th}Sentence of the source text.
$\boldsymbol{T_j}$**:** Thej^{th}Sentence of the target text.
$\boldsymbol{d}$ **:** A function based on a probability model, which defines the cost of every model, by couple(P_i, T_j).

a. The substitution of x_1 with y_1Generate the cost $d(x_1, y_1 ; 0, 0)$.
b. The abolition of x_1Generate the cost $d(x_1, 0 ; 0, 0)$.
c. The insertion of y_1Generate the cost $d(0, y_1 ; 0, 0)$.
d. The contraction of x_1 and x_2 to y_1 Generate the cost $d(x_1, y_1 ; x_2, 0)$.
e. The expansion of x_1 to y_1 and y_2 Generate the cost $d(x_1, y_1 ; 0, y_2)$.
f. The fusion of x_1 and x_2 Corresponding to y_1 and y_2 Generate the cost $d(x_1, y_1 ; x_2, y_2)$.

$\boldsymbol{D(i,j)}$**:** A function which determines the score of pair(P_i, T_j) this function return the minimum of six cases of model:

$$D(i,j) = min \begin{cases} D(i-1, j-1) + d(x_1, y_1 ; 0, 0) \\ D(i-1, j) + d(x_1, 0 ; 0, 0) \\ D(i, j-1) + d(0, y_1 ; 0, 0) \\ D(i-2, j-1) + d(x_1, y_1 ; x_2, 0) \\ D(i-1, j-2) + d(x_1, y_1 ; 0, y_2) \\ D(i-2, j-2) + d(x_1, y_1 ; x_2, y_2) \end{cases}$$

3.2.3 Advantages and disadvantages:

The simplicity of calculation gives an advantage to these types of methods, but the obtained results are lacking precision when it is about complex problems.

3.3 Methods with improving the exploitation of lexical information

To improve the quality of the alignment, several methods use other tracks such as:

- Bilingual external dictionaries as source of lexical matching (Method of Kay)
- Introduction to concept of « cognates » used by Simard and al. 1992 [11]

According to David, ed. (2011) [12], cognates or " related words " are words which have an origin, a common etymology. The term can indicate words of the same language or the words of different languages.

Example of cognates:

Words : sûr in frensh, *segur*in occitan and catalan, *seguro* in espagnol and portugais, *sicuro* in italien are cognates,because they are all directly stemming from the Latin word *sēcūrus*.

Forgeries-friends:

It happens that two words stemming from the same root evolve into different meanings, they become then what we call forgeries-friends. For example, The French word "journée" and the English word "journey" result both from the old former French of the XI[th] century jornee, meaning "journey or day labor". Since then, the English term kept the idea of the journey, and French held the temporal sense, which we know.

3.3.1 Method of Simard

Simard and al. (1992) detected the weakness of the classic methods based on the correlation of length for the complex problems.

His proposal was about calculating the "cognacity" of the sentences, that is to say, that a pair of sentences, which are mutual translations, contains much more of cognates than a random pair of sentences.

Simard proposed an algorithm of detection of cognates, as follows:

S_1 and S_2 a pair of sentences.
– Creation of the lists $T_1$and $T_2$of words t of every sentence;
– Comparison of the elements of both lists. Let's take two candidates t_1 and $t_1$of lists of words respectively T_1 and T_2;
- Categorization of the elements of lists. T is a candidate for a pair of cognate, if he corresponds to one of the following categories:
1. t consists completely of letters and number and contains at least a number;
2. t exclusively consists of letters and contains at least four letters;
3. t is a character of simple punctuation .
- $t_1$are $t_2$are cognates if and only if
  1. Both belong to the category 1 or 3 and that they are completely identical;
  2. Both belong to the category 2 and they have their four first characters identical.

Therefore, in order to identify a pair of cognates, Simard defines a common maximal sub-chain as an initial sub-chain containing at least four letters.

3.3.2 Method of Debili and Sammouda

The main steps of the method of Debili and Sammouda (1992) [13] are :

- Search for translations of chains in the target language, and vice versa.
- Comparison and evaluation (N) chains susceptible to be aligned.

Calculation of the global note of the alignment.

$$N = \left[1 - \frac{|L(c_1) - L(c_2)|}{L(c_1) + L(c_2)}\right] . \sum_{i=1}^{n(t)} t^2$$

With :
$\boldsymbol{L(c)}$: Number of character in the chain c .
$\boldsymbol{n(t)}$: The number of common maximal sub-chains with the length t.

Let's take:
$\boldsymbol{N_{f-a}}$: The best mark obtained in the direction f towards a.
$\boldsymbol{N_{a-f}}$: The best mark obtained in the direction a towards f.

The global note is: $\boldsymbol{N_{global} = N_{f-a} + N_{a-f}}$

Example:

1. Consultation with dictionary.

The word« ministère » contains as translations:

– *agency*
– *crown*
– *department*
– *ministère*
– *office*
«*minister*» contains :
Et– minister
– pasteur
– secrétaire

2. Comparison of the French chain « ministère » Comparison of the English chain «*minister*». The best note is obtained with the word« minister » :

$$N_{f-a} = \left[1 - \frac{|9-8|}{9+8}\right].(6^2 + 2^2) = 37{,}647$$

3. Comparison of the English chain " minister " with each of the translations of the French chain "ministère". The best note is obtained with the word « *ministry* » :

$$N_{a-f} = \left[1 - \frac{|8-8|}{8+8}\right].(6^2 + 1^2) = 37$$

The global note is :
$$\boldsymbol{N_{global} = N_{f-a} + N_{a-f} = 74{,}647}$$

3.3.3 Advantages & disadvantages

The notion of cognates allowed these types of methods to improve simply without using the lexical information.

On the other hand, it is impossible to apply these methods to the languages which do not belong to the same family.

The idea of Debili and Sammouda to make the evaluation in both senses improve well the precision of alignment.

3.4 Combines methods

The approach to these types of methods is to combine several methods to increase the quality of the alignment.

We find several methods existing in the literature, namely: Langlais (Langlais, 1997 [14]; Langlais & El-Bèze, 1997 [15]) , Simard & Plamondon (1998) [3] and the technique of Kraif (1999, 2001) [16][17].

3.4.1 Method of Langlais

This method integrates a model which takes care of the constraints of the surface (length of sentences and frequency of every model of translation) and the linguistic constraints; his mechanism of alignment consists in :

- Reduce the space of research, to return the less complex problem.
- An alignment which remits in an algorithm of dynamic programming for the research of the optimal alignment by considering scores putting surface as a linguistic indication

3.4.2 Method of Simard and Plamondon

Simard and Plamondon (1998 [18]) proposed a method which consisted in combining(organizing) the strength of the methods based on the information of the characters and the precision of the methods based on the lexical information.

Their approach consists in realizing two steps:

- First step: the mapping bi-textual (realized by means of the program Jacal which tries to put in correspondence the isolated cognates).
- Intermediate step: segmentation of the space of search.
- The second stage: alignment of the sentences (realized thanks to the program Salign based on the Model 1 of Brown).

3.4.3 Advantages and disadvantages:

The concept of a segment of research space improve the precision in a significant way, nevertheless, the problem bound to the notion of cognate limits its application in a set of languages restricts.

4 Application

In this section, we are going to use a tool to align two documents (one in French and the other one in amazighe)

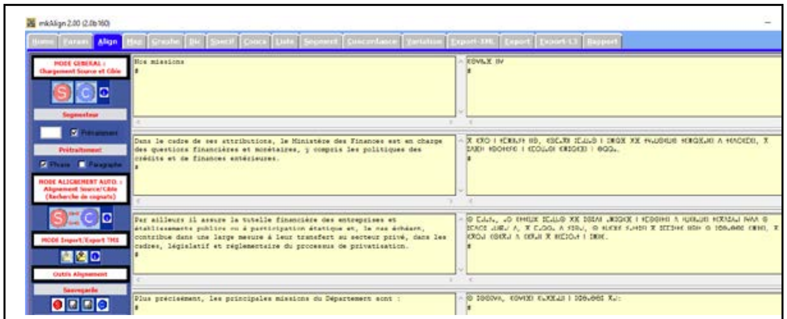


Fig. 2. Alignment

We take the example of two words « *Lancement* » having a frequency of 26 and « ⵜⵇⵥⵎ »having a frequency of 29.

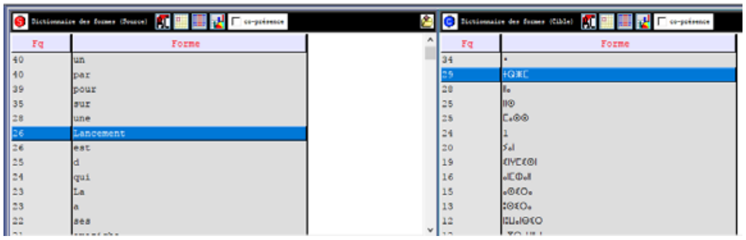


Fig. 3. Dictionary

The following figure presents the correlation between these two words:

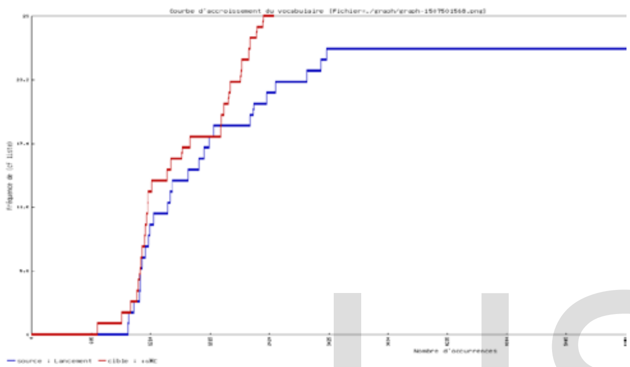


Fig. 4. Graph

Below *Map* mentioning the position of both words:

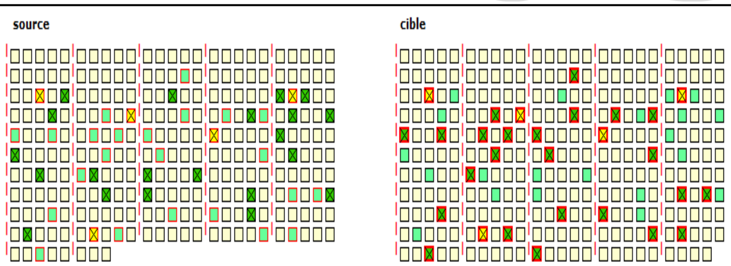


Fig. 5. Map

The concordance of both words:

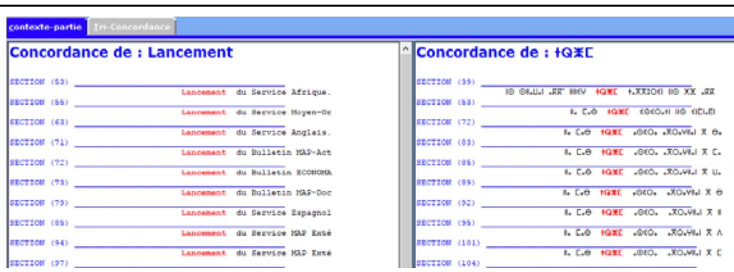


Fig. 6. Matching

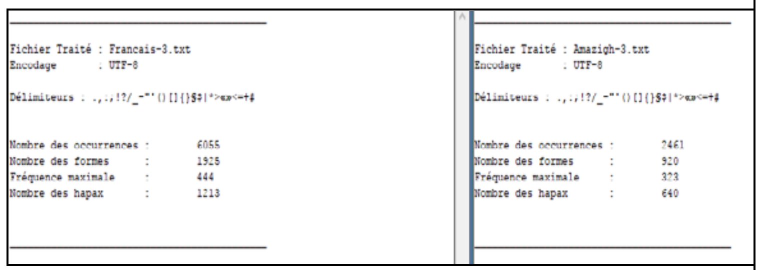


Fig. 7. Report

Conclusion

This paper describes the state of art on the methods used for the alignment of parallel texts. The characteristics of these methods have been mentioned in terms of advantages and drawbacks. It was found that the combined methods are precise and powerful when integrating both surface and lexical constraints.

The use of the combined approach enables an efficient alignment especially for languages belonging to different families (French-Amazighe).

The adoption of such an approach is going to allow us to realize an effective alignment especially for languages which do not belong in the same family (French language and the amazighe language).

References

[1] VÉRONIS, J. (2000).Parallel text processing: Alignment and use of translation corpora. Dordrecht: Kluwer Academic Publishers

[2] SIMARD, M. (1998). The BAF: A corpus of English-french bitext. In Proceedings of First International Conference on Language Resources and Evaluation (LREC),(pp. 489–496).

[3] BROWN, P. F., LAI, J. C., & MERCER, R. L. (1991). Aligning sentences in parallel corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, (pp. 169– 176).

[4] GALE, W. A., & CHURCH, K. W. (1993). A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(3), 75–102

[5] KAY, M., & RÖSCHEISEN, M. (1988). Text-translation alignment. Tech. rep., XeroxPalo Alto Research Center.

[6] KAY, M., & RÖSCHEISEN, M. (1993). Text-translation alignment. Computational Linguistics, 19(1), 121–142.

[7] KNUTH, D. E. (1997). The art of computer programming, vol. 3: Sorting and searching. Addison-Wesley, third ed

[8] VAN RIJSBERGEN, C. J. (1979). Information retrieval. Addison-Wesley, Butterworths ed.

[9] GALE, W. A., & CHURCH, K. W. (1991). Identifying word correspondences in parallel texts. In Proceedings of the 4th DARPA Speach and Natural Language Workshop, (pp. 152– 157).

[10] FUNG, P., & CHURCH, K. W. (1994). K-vec: A new approach for aligning parallel texts. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), vol. 2, (pp. 1096– 1102)

[11] SIMARD, M., FOSTER, G., & ISABELLE, P. (1992). Using cognates to align sentences in bilingual corpora. In Proceedings of the Fourth International Conference on theoretical and Methodological Issues in Machine Translation(TMI), (pp. 67 –81).

[12] David, ed. (2011). "cognate". A Dictionary of Linguistics and Phonetics (6th ed.). Blackwell Publishing. p. 104. (ISBN 978-1-4443-5675-5). OCLC 899159900. Retrieved 16 March 2016.

[13] DEBILI, F., & SAMMOUDA, E. (1992). Appariement des phrases de textes bilingues.In Proceedings of the 14th International Conference on Computational Linguistics (COLING 92), (pp. 517– 538)

[14] LANGLAIS, P. (1997). Alignement de corpus bilingues : intérêt, algorithmes et évaluation. In Bulletin de Linguistique Appliquée et Générale, numéro Hors Série, (pp. 245–254). Université de Franche-Comté.

[15] LANGLAIS, P., & EL-BÈZE, M. (1997). Alignement de corpus bilingues : algorithmes et évaluation. In Proceedings of 1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la langue de l'AUPELF-UREF (JST), (pp. 191–197).

[16] KRAIF, O. (1999). Identification des cognats et alignement bi-textuel : une étude empirique. In Actes de la 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 99), (pp. 205–214).

[17] KRAIF, O. (2001). Exploitation des cognats dans les systèmes d'alignement bitextuel : architecture et évaluation. TAL, 42(3)

[18] SIMARD, M., & PLAMONDON, P. (1998). Bilingual sentence alignment : Balancingrobustness and accuracy. Machine Translation, 13(1), 59–80.

[19] Mohamed Outahajala, Yassine Benajiba, Lahbib Zenkouar (2012). L'étiquetage grammatical de l'amazighe en utilisant les propriétés n-grammes et un prétraitement de segmentation (DOI 10.1007/978-3-642-22327-3_28).

[20] Mohamed Outahajala, Yassine Benajiba, Paolo Rosso, Lahbib Zenkouar (2011) POS Tagging in Amazighe Using Support Vector Machines and Conditional Random Fields.