A System to Align Complex Bilingual Corpora

Philippe Langlais

CTT-KTH – SE-10044, Stockholm, Sweden
CERI-LIA, Agroparc - BP 1228 - F-84911 Avignon Cedex 9, France
www.speech.kth.se

Abstract

Bilingual textual alignment systems are needed in many domains such as automatic or assisted translation, multi-lingual terminology and lexicography, multi-lingual information retrieval systems, etc. In the light of experiments carried out within the framework of a Research Concerted Action on multi-lingual alignment (namely ARCADE) initiated by AUPELF-UREF, we describe our bilingual alignment system which has proved to be efficient both for traditional corpora usually used to test such systems and for more complicated ones such as scientific articles or novels (Langlais et al., 1998). A comparative study of several functions which can be used to score the candidates for pairing as well as of several combinations of stages which are involved in an alignment system is made and discussed in the paper. It is important to mention that until now, most alignment systems have been evaluated on judicial and technical texts which present relatively few difficulties for a sentence-level alignement. However, other corpora such as novels are widespread and of utmost interest for many applications. In this respect, we report results on various English-French corpora (of several levels of difficulty) that have been made available within ARCADE. The paper shows that when aligning complex corpora, systems performances fall significantly, thus justifying the system we propose.

1 Introduction

In the last few years, there has been a growing interest in parallel text alignment techniques. These techniques attempt to map various textual units to their translation, and have proven useful for a wide range of applications (memory-based translation, extraction of multilingual lexical and terminological resources, semantic disambiguation, etc.) (Brown et al., 1991; Gale and Church, 1991a; Debili, 1992; Débili et al., 1994; Kay and Röscheisen, 1993; Simard et al., 1992; Simard and Plamondon, 1996).

A number of methods have been described in the literature and encouraging results have been reported (Gale and Church, 1991a; Simard and Plamondon, 1996). Unfortunately performance tends to deteriorate significantly when the tools are applied to corpora which are widely different from the training corpus, and/or where the alignments are not straightforward (for instance, graphics, tables, "floating" notes and missing segments, which are very common in real texts, and all of which result in a dramatic loss of efficiency).

1.1 A Brief overview of the ARCADE exercice

ARCADE, is an evaluation exercise financed by AUPELF-UREF, a network of (at least partially) French-speaking universities. It was launched in 1995 in order to promote research in the field of multi-lingual alignment. The first 2-year period (96-97) was dedicated to two main tasks: 1) the production of a reference bilingual corpus (French-English) aligned at sentence level; 2) the evaluation of several sentence alignment systems though an ARPA-like competition.

In its first phase, ARCADE was organized around two types of teams: the corpus providers (LPL and RALI) and the participants in the competition (RALI, LORIA, ISSCO, IRMC and LIA). General coordination was handled by J. Véronis (LPL); a discussion group was set up, and was moderated by Ph. Langlais (LIA & KTH).

2 Description of a new system: JAPA

As many systems, JAPA involves three major steps that are fully described: 1) the selection of potential pairs of sentences, 2) the scoring

of each of these pairs and 3) the selection of the optimal alignment according to a scoring function.

2.1 The scoring function

JAPA makes use of information that has been investigated in other studies, but integrates them in a convenient and efficient way.

2.1.1 Non-linguistic information

One of the earliest information that has been used to align texts is the length of the segments to align. In this respect, two models have been proposed : the former considering the length of segments counted in characters (Gale and Church, 1991a), the latter considering the length counted in words (Brown et al., 1991). The underlying idea of both of these models is that the lengths of the translated segments are proportional. Gale & Church proposed a probabilistic model which produces an approximation of the probability that two segments are mutual translation, given their lengths and the likelihood of the translation pattern that connect them :

$S_{lg} = -log\,[Prob(\delta|match) \times P(match)]$ where

δ is computed directly from the length of the segments (see (Brown et al., 1991) for details on the estimation of $Prob(\delta|match)$) and $P(match)$ is the a priori probability of the considered translation pattern. A pattern is just defined by the number of sentences that are considered both in the source and the target version (ex: 1-3 means that one source sentence is associated with 3 target ones).

It is interesting to note that such a simple model has been proved to give good results on huge corpora. It can however be argued that the corpora aligned with this model were mostly easy ones (Simard et al., 1992). The section 3 illustrates this point.

2.1.2 Lexical information

A more intuitive idea when a human confronts with the problem of aligning a corpus (even when he does not know perfectly the languages under consideration) is to use information conveyed by words. It is well known that for historical reasons, many languages (but not all) share many words or at least *lemmata*. This is particularly true if the languages under consideration are European ones. Thus, a natural way to achieve alignment is to use a bilingual lexicon. Unfortunately, such open and free lexicons are not widespread over our community. We explored two alternatives to get around these problems. The former is the automatic extraction of a bilingual lexicon in an incremental way; the latter is the integration of the so called cognates (Simard et al., 1992).

Extraction of a bilingual lexicon. JAPA uses as an option a set of bilingual lexicons. Some of them are coming from Internet (thematic lexicons) and others are issued from an automatic extraction process which makes use of sentence-aligned bilingual corpora. Our lexicon has a total of around 12000 entries. To this extent we used the likelihood test (tv) which is quiet simple and which behaviour has been judged satisfactorily in previous studies (Dunning, 1993; Gaussier and Langé, 95).

$$\begin{aligned} tv \;=\; & a\,log\,a + b\,log\,b + c\,log\,c + d\,log\,d \\ & -(a+c)\,log(a+c) - (a+b)\,log(a+b) \\ & -(b+d)\,log(b+d) - (c+d)\,log(c+d) \\ & +N\,log\,N \end{aligned}$$

where a stands for the number of areas wherein both e and f are observed ; b and c stand for the number of times where only one of the two words is encountered (resp. e and f) ; d is the number of the area where none of the two words are present and N stands for the total number of areas in the corpus.

Lists of a maximum of ten candidates to the translation of each considered word (mostly plain-words) have been selected with this metric and filtered by imposing constraints such as reciprocity (Gaussier and Langé, 95) and threshold-like rules. These heuristics are not fully satisfactory since many word-correspondences involve not only words but also terms (especially in domain-specific vocabulary). It is however a good compromise to arrive at a useful bilingual lexicon.

As we use word correspondences in a static way, we can mention the alignment system proposed by Kay and Röscheisen (1993) that uses a similar measurement in a dynamic process. Blank (1995) discusses the advantages and disadvantages of such a system.

Cognates. Simard et al. (1992) proposed a measurement of the bond of two segments, based on the notion of cognate that is defined as a pair of words (one word for each language) which share obvious properties at whatever level (phonologic, orthographic, semantic,...). Pairs such as accès/access, activité/activity are examples of French-English cognates. This definition can also be extended to entities which are not modified much during translation, such as proper nouns, numerical data, or also some punctuation marks. The authors proposed few rules to automatically select cognates in bilingual corpora: a) two words which are composed by at least one digit are cognates if they are identical, b) same punctuation marks are cognates, and last but not least, c) two alpha-words (composed of letters only) are cognates if they shared the same n-first characters. The authors proposed to score a candidate for pairing by S_{cog} :

$$S_{cog} = \frac{P_T(c|n)}{P_R(c|n)} \text{ , where}$$

$P_T(c|n)$ (resp. $P_R(c|n)$) is the probability that a source-segment of n words shares c cognates with his target-counterpart under the hypothesis that they are mutual translation (resp. are selected randomly). Both this probabilities follow approximately a binomial distribution where p_T (resp p_R) is the probability that a source-word is part of a cognate when segments that are mutual translation (resp. are randomly selected). p_T and p_R have been experimentally set up to 0.3 and 0.09.

$$\begin{aligned} P_T(c|n) &= C_n^p \times p_T^c \times (1-p_T)^{n-c} \\ P_R(c|n) &= C_n^p \times p_R^c \times (1-p_R)^{n-c} \end{aligned}$$

It is interesting to note that the results obtained by this approach have been reported to be less accurate than the ones reported by Gale and Church (1991a).

2.1.3 The final score

JAPAs scoring function uses the information we described. Its origin is the extension of the score proposed by (Simard et al., 1992) to the following one :

$$S_c = -log\left[\frac{P_T(c|n)}{P_R(c|n)} \times P(\delta|match) \times P(match)\right]$$

Once developed, it becomes the following, involving three quantities (x, y and z) which stand respectively for the cognate-score, the length-score and the pattern-score.

$$\begin{aligned} S_c \;=\; & -\left[c.log\tfrac{p_T}{p_R}\right] - \left[(n-c).log\tfrac{1-p_T}{1-p_R}\right] && (x) \\ & -log\, P(\delta|match) && (y) \\ & -log\, P(match) && (z) \end{aligned}$$

We can observe in table 1 that these quantities do not have the same dynamics. In a first attempt, we tried to normalize each quantity (using their z-score) with a significant loss of accuracy of the system ; thus, leading us to the conclusion, that each information is not of equal importance in the alignment process. According to this observation, we decided to find three ponderation coefficients (α_x, α_y and α_z) in order to weight each source of information; leading to the score expressed by S_{japa}. Note that this score is no longer a probability function and that it makes the assumption (not fully satisfied) that the different scores weigthed are statistically independant. We used the non-derivative minimization-technique called Simplex (Nedel and Mead, 1965) to find the combination which optimize the performance (both precision and recall) of the system on a corpus of 1000 hand-labeled pairs. The following values (possibly a local optima) have been found and are presently used in JAPA : $< \alpha_x = 0.5$, $\alpha_y = 0.2$,$\alpha_z = 1 >$.

$$\begin{aligned} S_{japa} \;=\; & -0.5 \times \left[c.log\tfrac{p_T}{p_R} - (n-c).log\tfrac{1-p_T}{1-p_R}\right] \\ & -0.2 \times log\, P(\delta|match) \\ & -log\, P(match) \end{aligned}$$

Both the cognates (dynamically detected) and the entries of the bilingual lexicon are considered by the cognate-score. We verified on

score	μ	σ	min	max
x	0.2	9.4	-104.1	36.2
y	13	16	0.1	69.07
z	2.8	1.7	0	4.5

Table 1: Average (μ), standard deviation (σ) and dynamics of the three quantities used in JAPAs scoring function. These values have been measured on a corpus of 1000 hand-labeled pairs.

a test-corpus the assumption that the number of cognates (extended to the entries of the bilingual lexicon) of a pair of sentences still remains modelized by a binomial distribution.

2.2 Selection of candidates for pairing

Even if it can be considered as an implementation detail, search-space reduction is a step that needs to be carefully handled. Our space-reduction method is based on the idea that aligning sentences can be done efficiently using a word-level alignment. As underlined by Debili (1992), we are faced to a vicious circle from which we can exit considering that a fine sentence-level alignment can use a coarse word-level one.

This is a solution that has also been implemented in (Simard et al., 1992). The authors proposed an algorithm to determine points where the solution has to pass through. This solution is efficient as far as the location of the points is accurate, which is a tricky point. We describe a space-reduction methodology which is less directive, but still remains efficient.

A bilingual corpus is represented as a binary matrix M, where the i_{th} line stands for the i_{th} word of the source text and the column j stands for the j_{th} word of the target text. The cell $M(i,j)$ is set to 1 if the source word i is in *relation* with the target word j. That is, the two words are either cognates, or are one of the entry of the bilingual lexicon.

An example of such a matrix for an extract from the novel of Jules Verne *"De la terre à la lune"* (the entire novel is referenced VERNE hereafter) is given in figure 1. We observe that

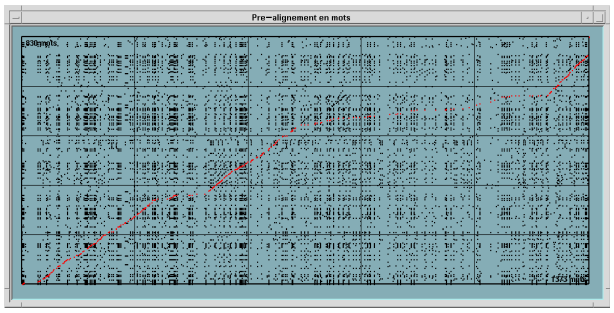

Figure 1: Binary matrix of words computed for an extract of a novel of Jules Verne (1373 source-words $\times$ 830 target-words). A dot indicates a relation between two words. Each word of the extract are considered in this example.

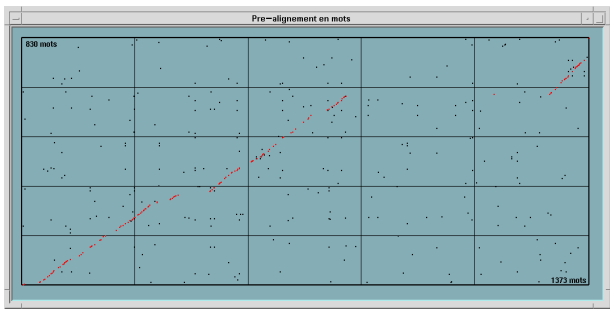

Figure 2: Binary matrix computed considering only words appearing less than 10 times.

a lot of source-words are in relation with many target-words (black columns), which is characteristic of over-represented tools-words.

To remove these "noisy" pairs of words, we can apply image-filtering techniques (as it has been proposed for example in (Chang and Chen, 1997)) or more simply, we can focus only on low-frequency words. The figure 2 shows the binary matrix computed on the same excerpt from VERNE, restricted to those words that appear less than 10 times.

We make the assumption that the pairing of low-frequency words is nearly a synchronized process that can be handled using a dynamic programming scheme. We define a cost for pairing two words (S) as:

$$S(I,J) = \min_{i=I-R}^{I-1} \min_{j/M_{i,j}=1} (S(i,j) + F(i,j,I,J))$$

$$\text{with } F(i,j,I,J) = \frac{J-j}{I-i} + (I-i-1) \times \mathcal{C}$$

where the contants $\mathcal{C}$ and R have been set up empirically from a training corpus. The underlying idea of this score is to minimize the deviation from the diagonal that should be observed if the assumption of the synchronization of low-frequency words is fully satisfied.

Thus, the search-beam at the sentence level is simply defined as a fixed-size number of sentences (here 8) centered around the word-alignment. This fast reduction method is accurate, as illustrated in section 3.4.

2.3 Strategy of sentence-alignment

We also make use of a dynamic programming scheme to align the sequence $(s_1, \ldots, s_I)$ of source-sentences with the sequence $(t_1, \ldots, t_J)$ of target-ones. The algorithm follows the one given by (Gale and Church, 1991a)

Only pairs belonging to the beam-search are taken into account in this process.

3 Experiments

3.1 Description of the corpora

These experiments have been carried out using the two French-English corpora available in the ARCADE framework : BAF and JOC. Both corpora have been aligned at the sentence level, and manually checked. JOC gathers 10 homogeneous institutional texts for a total size of 4 megabytes ; about 9000 pairs of sentences, most of them (94%) being 1 to 1 ones. BAF is a mix of 11 texts from various origin such as scientific articles, excerpts from the hansard-parliaments texts and novels. The total size of BAF is 6 megabytes and represents above 22000 pairs of sentences; 90% of them are 1-1 ones. Some of these texts have been judged difficult to align. Especially the VERNE corpus which is very interesting because the translations are sometimes divergent (75% of 1-1 patterns) and it is not even clear whether the English version is really a translation of the French one, or if it has been translated from an abridged version.

3.2 Tests

In order to better understand the efficiency of JAPA, we set up several systems that are described in table 2. They all share the same structure but differ either by the scoring function they use, either by the fact that word-alignment is performed or not. Two of these systems – GC and SI – are implementations of systems previously described in the literature (resp. (Gale and Church, 1991a) and (Simard et al., 1992)). For comparison purposes, we also report results given in Simard and Plamondon (1996).

	J	A	B	C	GC	SI	F	G
word	×		×	×			×	×
pond.	×	×	×					×
length	×	×	×	×	×			
cognate	×	×				×	×	×

Table 2: Description of the tested systems. **word** indicates that word-alignment is performed, **pond.** that S_{japa} is used to score the sentence-pairing, **length** that S_{lg} is used, and **cognate** that S_{cog} is used in the scoring function.

3.3 Evaluation

The quality of an alignment A is assessed via two rates as defined in (Simard and Plamondon, 1996) : the precision rate ($P = |A \cap A_{ref}|/|A|$) and a the recall rate ($R = |A \cap A_{ref}|/|A_{ref}|$); where A_{ref} stands for the reference (manually checked). Following the authors, we report these two rates computed at the character level rather than at the sentence one ; thus taking into account the size (counted in characters) of the alignment errors.

3.4 Results

Precision and recall rates computed at the character level are reported in table 3 both for BAF and JOC. The results observed on the VERNE are also reported to analyse the behaviour of the different systems on this particularly complex corpus.

First of all we can see that JAPA outperforms other systems both on BAF and JOC. It even outperforms Sa (Simard and Plamondon, 1996) which makes use of a statistical-translation model in its scoring function. Secondly, we can check that weighting the infor-

mation used to score an alignment is efficient for both corpora (JAPA vs. A). Fortunately, the word-alignment stage is also fruitful (SI vs F, GC vs C) especially on the BAF corpus. We can also observe a difference between results on BAF and JOC ; where BAF presents a more challenging corpus to align. In particular we can observe that GC is able to align only half of VERNE with a precision close to 0.4. In accordance with Simard et al. (1992) observations, GC outperforms slightly SI on the JOC corpus ; it is not true any longer when considering the rates on BAF. It is at this point interesting to note that the system G (which makes no use of the length-score $P(\delta|match)$ as defined by Gale and Church (1991b)) is still accurate both on BAf and JOC.

	BAF (P, R)	JOC (P, R)	VERNE (P, R)
J	(97.6,83.2)	(98.6,98.9)	(90.4,93.8)
A	(87.2,81.6)	(85.9,99.0)	(78.9,93.0)
B	(91.8,78.3)	(98.4,97.6)	(51.0,58.8)
C	(91.7,79.2)	(97.9,97.8)	(52.0,62.3)
GC	(60.7,47.8)	(98.0,97.9)	(41.4,50.4)
SI	(64.6,52.8)	(94.1,98.8)	(84.2,93.9)
F	(93.7,83.1)	(94.0,98.8)	(84.1,93.8)
G	(96.7,82.7)	(98.8,98.6)	(83.3,91.2)
Sa	(92.3,93.9)	—	(54.5,94.0)

Table 3: Alignment results. Precision and recall are given in percentage both for BAf and JOC, but also for VERNE which is a "non-easy" corpus. Average values are weighted by text size. For comparison purposes, Sa indicates the rates reported in (Simard and Plamondon, 1996).

The final ranking off the systems tested within the ARCADE exercice (including the Japa system) on both JOC and BAF corpus is also given in Figure 3. The global efficiency of the different systems are given as *F-measure* (Rijsbergen, 1979) which combines recall and precision in a single efficiency measure (harmonic mean of precision and recall): $F = 2.(recall \times precision)/(recall + precision)$. Recall and precision rates were computed here both at the sentence level and at the character

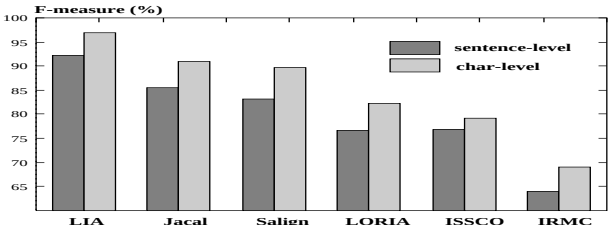

Figure 3: Final ranking off the systems (average F-values).

level. This last measure has been proposed within the ARCADE exercice in order to take into account the fact that alignment errors involving short sentences should be less penalized than errors involving longer ones, at least from the perspective of some applications. Refer to (Langlais et al., 1998) for further details on the evaluation protocol used in ARCADE and for a description of the different systems tested. As it can be observed, Japa outperforms other systems.

4 Conclusions

We have described a new bilingual alignment system which aligns sentences using first a word-alignment stage. Compared to previous systems described in the literature, JAPAs performances are fairly stable and very good, whatever the level of difficulty of the corpus to align. This study also confirms the importance of the choice of test-corpora in an evaluation stage and also shows that aligning bilingual corpora, even at the sentence level is not yet a solved problem.

References

I. Blank. 1995. Sentence alignment : Methods and implementations. *T.A.L.*, 36(1-2):81–99.

P.F. Brown, J.C. Lai, and R.L. Mercer. 1991. Aligning Sentences in Parallel Corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley,CA,USA.

J.S. Chang and M.H. Chen. 1997. An alignment method for noisy parallel corpora based on image processing techniques. In *Proceedings of 35th Meeting of the Association for Computational Linguistics*, pages 297–304, Madrid.

F. Débili, E. Sammouda, and A. Zribi. 1994. De l'appariement des mots à la comparaison de phrases. In *9ème Congrès de Reconnaissance des Formes et Intelligence Artificielle*, Paris, Janvier.

F. Debili. 1992. Aligning Sentences in Bilingual Texts French - English and French - Arabic. In *COLING*, pages 517–525, Nantes, 23–28 Aout.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19,1.

W. A. Gale and Kenneth W. Church. 1991a. A Program for Aligning Sentences in Bilingual Corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA.

W.A. Gale and K.W. Church. 1991b. Identifying word correspondences in parallel texts. In *Fourth DARPA Speech and Natural Language Workshop*, pages 152–157, Pacific Grove, CA, USA.

É. Gaussier and J-M. Langé. 95. Modèles statistiques pour l'extraction de lexiques bilingues. *T.A.L.*, 36(1-2):133–155.

M. Kay and M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Ph. Langlais, M. Simard, J. Véronis, S. Armstrong, P. Bonhomme, F. Débili, P. Isabelle, E. Souissi, and P. Théron. 1998. Arcade: A cooperative research project on parallel text alignment evaluation. In *First International Conference on Language Resources and Evaluation*, Granada, Spain.

J.A. Nedel and R. Mead. 1965. A simplex method for function minimization. *Comput. J.*, pages 308–313.

C.J. Van Rijsbergen. 1979. Information Retrieval,2nd edition, London, Butterworths.

M. Simard and P. Plamondon. 1996. Bilingual sentence alignment: Balancing robustness and accuracy. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA)*, Montréal, Québec.

M. Simard, G.F. Foster, and P. Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 67–81, Montréal, Canada.