Methods and Practical Issues in Evaluating Alignment Techniques

Philippe Langlais
CTT/KTH SE-10044 Stockholm
CERI-LIA, AGROPARC BP 1228
F-84911 Avignon Cedex 9
Philippe.Langlais@speech.kth.se

Michel Simard
RALI-DIRO
Univ. de Montréal
Québec, Canada H3C 3J7
simardm@IRO.UMontreal.CA

Jean Véronis
LPL, Univ. de Provence
29, Av. R. Schuman
F-13621 Aix-en-Provence Cedex 1
veronis@univ-aix.fr

Abstract

This paper describes the work achieved in the first half of a 4-year cooperative research project (ARCADE), financed by AUPELF-UREF. The project is devoted to the evaluation of parallel text alignment techniques. In its first period ARCADE ran a competition between six systems on a sentence-to-sentence alignment task which yielded two main types of results. First, a large reference bilingual corpus comprising of texts of different genres was created, each presenting various degrees of difficulty with respect to the alignment task.

Second, significant methodological progress was made both on the evaluation protocols and metrics, and the algorithms used by the different systems. For the second phase, which is now underway, ARCADE has been opened to a larger number of teams who will tackle the problem of word-level alignment.

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 and tools. A simple example of such a tool is probably the *TransSearch* bilingual concordancing system (Isabelle et al., 1993), which allows a user to query a large archive of existing translations in order to find ready-made solutions to specific translation problems. Such a tool has proved extremely useful not only for translators, but also for bilingual lexicographers (Langlois, 1996) and terminologists (Dagan and Church, 1994). More sophisticated applications based on alignment technology have also been the object of recent work, such as the automatic building of bilingual lexical resources (Melamed, 1996; Klavans and Tzoukermann, 1995), the automatic verification of translations (Macklovitch, 1995), the automatic dictation of translations (Brousseau et al., 1995) and even interactive machine translation (Foster et al., 1997).

Enthusiasm for this relatively new field was sparked early on by the apparent demonstration that very simple techniques could yield almost perfect results. For instance, to produce sentence alignments, Brown et al. (1991) and Gale and Church (1991) both proposed methods that completely ignored the lexical content of the texts and both reported accuracy levels exceeding 98%. Unfortunately performance tends to deteriorate significantly when aligners 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, all result in a dramatic loss of efficiency.

The truth is that, while text alignment is mostly an easy problem, especially when considered at the sentence level, there are situations where even humans have a hard time making the right decision. In fact, it could be argued that, ultimately, text alignment is no easier than the more general problem of natural language understanding.

In addition, most research efforts were directed towards the easiest problem, that of sentence-to-sentence alignment (Brown et al., 1991; Gale and Church, 1991; Debili, 1992; Kay and Röscheisen, 1993; Simard et al., 1992; Simard and Plamondon, 1996). Alignment at the word and term level, which is extremely useful for applications such as lexical resource extraction, is still a largely unexplored research area(Melamed, 1997).

In order to live up to the expectations of the

various application fields, alignment technology will therefore have to improve substantially. As was the case with several other language processing techniques (such as information retrieval, document understanding or speech recognition), it is likely that a systematic evaluation will enable such improvements. However, before the ARCADE project started, no formal evaluation exercise was underway; and worse still, there was no multilingual aligned reference corpus to serve as a "gold standard" (as the Brown corpus did, for example, for part of speech tagging), nor any established methodology for the evaluation of alignment systems.

2 Organization

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

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

3 Reference corpus

One of the main results of ARCADE has been to produce an aligned French-English corpus, combining texts of different genres and various degrees of difficulty for the alignment task. It is important to mention that until ARCADE, most alignment systems had been tested on judicial and technical texts which present relatively few difficulties for a sentence-level alignment. Therefore, diversity in the nature of the texts was preferred to the collection of a large quantity of similar data.

3.1 Format

ARCADE contributed to the development and testing of the Corpus Encoding Standard (CES), which was initiated during the MULTEXT project (Ide et al., 1995). The CES is based on SGML and it is an extension of the now internationally-accepted recommendations of the Text Encoding Initiative (Ide and Véronis, 1995). Both the JOC and BAF parts of the ARCADE corpus (described below) are encoded in CES format.

3.2 JOC

The JOC corpus contains texts which were published in 1993 as a section of the C Series of the Official Journal of the European Community in all of its official languages. This corpus, which was collected and prepared during the MLCC and MULTEXT projects, contains, in 9 parallel versions, questions asked by members of the European Parliament on a variety of topics and the corresponding answers from the European Commission. JOC contains approximately 10 million words (ca. 1.1 million words per language). The part used for JOC was composed of one fifth of the French and English sections (ca. 200 000 words per language).

3.3 BAF

The BAF corpus is also a set of parallel French-English texts of about 400 000 words per language. It includes four text genres: 1) **INST**, four institutional texts (including transcription of speech from the Hansard corpus) for a totaling close to 300 000 words per language, 2) **SCIENCE**, five scientific articles of about 50 000 words per language, 3) **TECH**, technical documentation of about 40 000 words per language and 4) **VERNE**, the Jules Verne novel: *"De la terre à la lune"* (ca. 50 000 words per language). This last text is very interesting because the translation of literary texts is much freer than that of other types of tests. Furthermore, the English version is slightly abridged, which adds the problem of detecting missing segments. The BAF corpus is described in greater detail in (Simard, 1998).

4 Evaluation measures

We first propose a formal definition of parallel text alignment, as defined in (Isabelle and Simard, 1996). Based on that definition, the usual notions of recall and precision can be used to evaluate the quality of a given alignment with

respect to a reference. However, recall and precision can be computed for various levels of granularity: an alignment at a given level (e.g. sentences) can be measured in terms of units of a lower level (e.g. words, characters). Such a fine-grained measure is less sensitive to segmentation problems, and can be used to weight errors according to the number of sub-units they span.

4.1 Formal definition

If we consider a text S and its translation T as two sets of segments $S = \{s_1, s_2, .., s_n\}$ and $T = \{t_1, t_2, ..., t_m\}$, an *alignment* A between S and T can be defined as a subset of the Cartesian product $\wp(S) \times \wp(T)$, where $\wp(S)$ and $\wp(T)$ are respectively the set of all subsets of S and T. The triple (S, T, A) will be called a *bitext*. Each of the elements (ordered pairs) of the alignment will be called a *bisegment*.

This definition is fairly general. However, in the evaluation exercice described here, segments were sentences and were supposed to be contiguous, yielding *monotonic alignments*.

For instance, let us consider the following alignment, which will serve as the *reference alignment* in the subsequent examples. The formal representation of it is: $A_r = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_2, t_3\})\}$.

s_1 Phrase numéro un.	t_1 The first sentence.
s_2 Phrase numéro deux qui ressemble à la 1ère.	t_2 The 2nd sentence. t_3 It looks like the first.

4.2 Recall and precision

Let us consider a bitext (S, T, A_r) and a proposed alignment A. The alignment *recall* with respect to the reference A_r is defined as: $recall = |A \cap A_r|/|A_r|$. It represents the proportion of bisegments in A that are correct with respect to the reference A_r. The *silence* corresponds to $1 - recall$. The alignment *precision* with respect to the reference A_r is defined as: $precision = |A \cap A_r|/|A|$. It represents the proportion of bisegments in A that are right with respect to the number of bisegment proposed. The *noise* corresponds to $1 - precision$.

We will also use the *F-measure* (Rijsbergen, 1979) which combines recall and precision in a single efficiency measure (harmonic mean of precision and recall):

$$F = 2.\frac{(recall \times precision)}{(recall + precision)}.$$

Let us assume the following proposed alignment:

s_1 Phrase numéro un.	t_1 The first sentence.
	t_2 The 2nd sentence.
s_2 Phrase numéro deux qui ressemble à la 1ère.	t_3 It looks like the first.

The formal representation of this alignment is: $A = \{(\{s_1\}, \{t_1\}), (\{\}, \{t_2\}), (\{s_2\}, \{t_3\})\}$. We note that: $A \cap A_r = \{(\{s_1\}, \{t_1\})\}$. Alignment recall and precision with respect to A_r are $1/2 = 0.50$ and $1/3 = 0.33$ respectively. The F-measure is 0.40.

Improving both recall and precision are antagonistic goals : efforts to improve one often result in degrading the other. Depending on the applications, different trade-offs can be sought. For example, if the bisegments are used to automatically generate a bilingual dictionary, maximizing precision (i.e. omitting doubtful couples) is likely to be the preferred option.

Recall and precision as defined above are rather unforgiving. They do not take into account the fact that some bisegments could be partially correct. In the previous example, the bisegment $(\{s_2\}, \{t_3\})$ does not belong to the reference, but can be considered as partially correct: t_3 does match a part of s_2. To take partial correctness into account, we need to compute recall and precision at the sentence level instead of the alignment level.

Assuming the alignment $A = \{a_1, a_2, \ldots, a_m\}$ and the reference $A_r = \{ar_1, ar_2, \ldots, ar_n\}$, with $a_i = (as_i, at_i)$ and $ar_j = (ars_j, art_j)$, we can derive the following sentence-to-sentence alignments:

$$A' = \bigcup_i (as_i \times at_i)$$
$$A'_r = \bigcup_j (ars_j \times art_j)$$

Sentence-level recall and precision can thus be defined in the following way:

$$\begin{aligned} recall &= |A' \cap A'_r|/|A'_r| \\ precision &= |A' \cap A'_r|/|A'| \end{aligned}$$

In the example above: $A' = \{(s1, t1), (s2, t3)\}$ and $A'_r = \{(s1, t1), (s2, t2), (s2, t3)\}$. Sentence-level recall and precision for this example are

therefore $2/3 = 0.66$ and 1 respectively, as compared to the alignment-level recall and precision, 0.50 and 0.33 respectively. The F-measure becomes 0.80 instead of 0.40.

4.3 Granularity

In the definitions above, the sentence is the unit of granularity used for the computation of recall and precision at both levels. This results in two difficulties. First, the measures are very sensitive to sentence segmentation errors. Secondly, they do not reflect the seriousness of misalignments. It seems reasonable that errors involving short sentences should be less penalized than errors involving longer ones, at least from the perspective of some applications.

These problems can be avoided by taking advantage of the fact that a unit of a given granularity (e.g. sentence) can always be seen as a (possibly discontinuous) sequence of units of finer granularity (e.g. character).

Thus, when an alignment A is compared to a reference alignment A_r using the recall and precision measures computed at the char-level, the values obtained are inversely proportional to the quantity of text (i.e. number of characters) in the misaligned sentences, instead of the number of these misaligned sentences. For instance, in the example used above, we would have at sentence level:

- using word granularity (punctuation marks are considered as words) :

```
|A'| = 4*4 + 0*4 + 9*6 = 106
|Ar'| = 4*4 + 9*10 = 70
|Ar' ^ A'|  = 4*4 + 9*6 = 70
recall = 70/106 = 0.66
precision = 1
F = 0.80
```

- using character granularity (excluding spaces):

```
|A'| = 15*17 + 0*15 + 36*20 = 975
|Ar'| = 15*17 + 36*35 = 1515
|Ar' ^ A'| = 15*17 + 36*20 = 975
recall = 975/1515 = 0.64
precision = 1
F = 0.78
```

5 Systems tested

Six systems were tested, two of which having been submitted by the RALI.

RALI/Jacal This system uses as a first step a program that reduces the search space only to those sentence pairs that are potentially interesting (Simard and Plamondon, 1996). The underlying principle is the automatic detection of isolated cognates (i.e. for which no other similar word exists in a window of given size). Once the search space is reduced, the system aligns the sentences using the well-known sentence-length model described in (Gale and Church, 1991).

RALI/Salign The second method proposed by RALI is based on a dynamic programming scheme which uses a score function derived from a translation model similar to that of (Brown et al., 1990). The search space is reduced to a beam of fixed width around the diagonal (which would represent the alignment if the two texts were perfectly synchronized).

LORIA The strategy adopted in this system differs from that of the other systems since sentence alignment is performed after the preliminary alignment of larger units (whenever possible, using mark-up), such as paragraphs and divisions, on the basis of the SGML structure. A dynamic programming scheme is applied to all alignment levels in successive steps.

IRMC This system involves a preliminary, rough word alignment step which uses a transfer dictionary and a measure of the proximity of words (Débili et al., 1994). Sentence alignment is then achieved by an algorithm which optimizes several criteria such as word-order conservation and synchronization between the two texts.

LIA Like Jacal, the LIA system uses a pre-processing step involving cognate recognition which restricts the search space, but in a less restrictive way. Sentence alignment is then achieved through dynamic programming, using a score function which combines sentence length, cognates, transfer dictionary and frequency of translation schemes (1-1, 1-2, etc.).

ISSCO Like the LORIA system, the ISSCO aligner is sensitive to the macro-structure of the document. It examines the tree structure of an SGML document in a first pass, weighting each node according to the number of characters contained within the subtree rooted at that node. The second pass descends the tree, first

by depth, then by breath, while aligning sentences using a method resembling that of Gale & Church.

6 Results

Four sets of recall/precision measures were computed for the alignments achieved by the six systems for each text type previously described above: **Align**, alignment-level, **Sent** sentence-level, **Word**, word-level and **Char**, character-level. The global efficiency of the different systems (average F-values) for each text type is given in Figure 1.

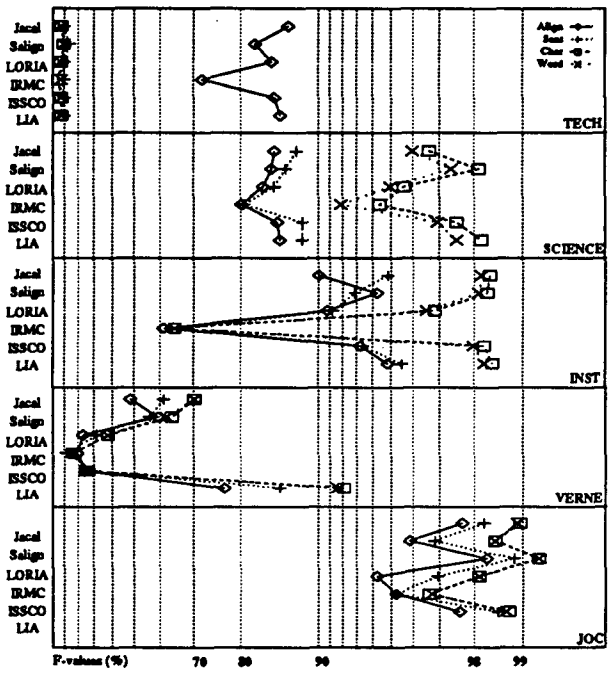

Figure 1: Global efficiency (average F-values for **Align**, **Sent**, **Word** and **Char** measures) of the different systems (Jacal, Salign, LORIA, IRMC, ISSCO, LIA) , by text type (logarithmic scale).

First, note than the *Char* measures are higher that the *Align* measures. This seems to confirm that systems tend to fail when dealing with shorter sentences. In addition, the reference alignment for the BAF corpus combines several 1-1 alignments in a single n-n alignment, for practical reasons owing to the sentence segmentation process. This results in decreased *Align* measures.

The corpus on which all systems scored highest was the JOC. This corpus is relatively simple to align, since it contains 94% of 1-1 alignments, reflecting a translation strategy based on speed and absolute fidelity. In addition, this corpus contains a large amount of data that remains unchanged during the translation process (proper names, dates, etc.) and which can serve as anchor points by some systems. Note that the LORIA system achieves a slightly better performance than the others on this corpus, mainly because it is able to carry out a structure-alignment since paragraphs and divisions are explicitly marked.

The worst results were achieved on the VERNE corpus. This is also the corpus for which the results showed the most scattering across systems (22% to 90% char-precision). These poor results are linked to the literary nature of the corpus, where translation is freer and more interpretative. In addition, since the English version is slightly abridged, the occasional omissions result in de-synchronization in most systems. Nevertheless, the LIA system still achieves a satisfactory performance (90% char-recall and 94% char-precision), which can be explained by the efficiency of its word-based pre-alignment step, as well as the scoring function used to rank the candidate bisegments.

Significant discrepancy are also noted between the *Align* and *Char* recalls on the TECH corpus. This document contained a large glossary as an appendix, and since the terms are sorted in alphabetic order, they are ordered differently in each language. This portion of text was not manually aligned in the reference. The size of this bisegment (250-250) drastically lowers the *Char*-recall. Aligning two glossaries can be seen as a document-structure alignment task rather than a sentence-alignment task. Since the goal of the evaluation was sentence alignment, the TECH corpus results were not taken into account in the final grading of the systems.

The overall ranking for all systems (excluding the TECH corpus results) is given in Figure 2, in terms of the *Sent* and *Char* F-measures. The LIA system obtains the best average results and shows good stability across texts, which is an

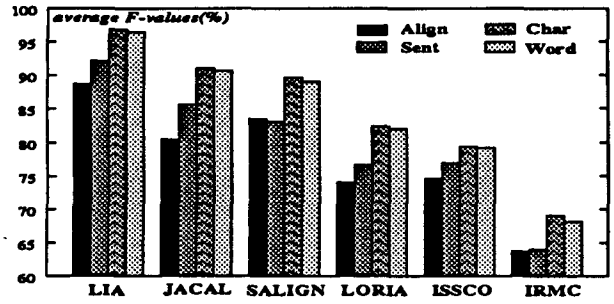

Figure 2: Final ranking on the systems (average F-values).

important criterion for many applications.

7 Conclusion and future work

The ARCADE evaluation exercise has allowed for significant methodological progress on parallel text alignment. The discussions among participants on the question of a testing protocol resulted in the definition of several evaluation measures and an assessment of their relative merits. The comparative study of the systems performance also yielded a better understanding of the various techniques involved. As a significant spin-off, the project has produced a large aligned bilingual corpus, composed of several types of texts, which can be used as a gold standard for future evaluation. Grounded on the experience gained in the first test campaign, the second (1998-1999) has been opened to more teams and plans to tackle more difficult problems, such as word-level alignment. [1].

Acknowledgments

This work has been partially funded by AUPELF-UREF. We are indebted to Lucie Langlois and Elliott Macklovitch for their fruitful comments on this paper.

References

J. Brousseau, C. Drouin, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P. Plamondon. 1995. French Speech Recognition in an Automatic Dictation System for Translators: the TransTalk Project. In *Proceedings of Eurospeech 95*, Madrid, Spain.

[1]For more information check the Web site at http://www.lpl.univ-aix.fr/projects/arcade

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roosin. 1990. A Statistical Approach to Machine Translation. In *Computational Linguistics*, volume 16, pages 79–85, June.

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.

Ido Dagan and Kenneth W. Church. 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of ANLP-94*, Stuttgart, Germany.

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.

George Foster, Pierre Isabelle, and Pierre Plamondon. 1997. Target-Text Mediated Interactive Machine Translation. *Machine Translation*, 21(1-2).

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

N. Ide and J. Véronis, 1995. *The Text Encoding Initiative: background and context*, chapter 342p. Kluwer Academic Publishers, Dordrecht.

N. Ide, G. Priest-Dorman, and J. Véronis. 1995. Corpus encoding standard. Report. Accessible on the World Wide Web: http://www.lpl.univ-aix.fr/projects/multext/CES/CES1.html.

Pierre Isabelle and Michel Simard. 1996. Propositions pour la représentation et l'évaluation des alignements de textes parallèles. `http://www-rali.iro.umontreal.ca/arc-a2/-PropEval`.

Pierre Isabelle, Marc Dymetman, George Foster, Jean-Marc Jutras, Elliott Macklovitch, François Perrault, Xiaobo Ren, and Michel

Simard. 1993. Translation Analysis and Translation Automation. In *Proceedings of TMI-93*, Kyoto, Japan.

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

Judith Klavans and Evelyne Tzoukermann. 1995. Combining Corpus and Machine-readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, 10(3).

Lucie Langlois. 1996. Bilingual Concordances: A New Tool for Bilingual Lexicographers. In *Proceedings of AMTA-96*, Montréal, Canada.

Elliott Macklovitch. 1995. TransCheck — or the Automatic Validation of Human Translations. In *Proceedings of the MT Summit V*, Luxembourg.

I. Dan Melamed. 1996. Automatic Construction of Clean Broad-coverage Translation Lexicons. In *Proceedings of AMTA-96*, Montréal, Canada.

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. In *35th Conference of the Association for Computational Linguistics*, Madrid, Spain.

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.

M. Simard. 1998. The BAF: A corpus of English-French Bitext. In *First International Conference on Language Resources and Evaluation*, Granada, Spain.