ARCADE: A Cooperative Research Project on Parallel Text Alignment Evaluation

Langlais Ph.(1,2), Simard M.(3), Véronis J.(4),
Armstrong S.(5), Bonhomme P.(6), Debili F.(7), Isabelle P.(3), Souissi E.(7), Théron P.(5)

(1) CTT-KTH – SE-10044, Stockholm, Sweden
(2)CERI-LIA, Agroparc - BP 1228 - F-84911 Avignon Cedex 9, France
(3) RALI, Université de Montréal, Québec, Canada
(4) LPL, Université de Provence & CNRS, – 29, Av. R. Schuman – F-13621 Aix-en-Provence Cedex 1, France
(5) ISSCO, Université de Genève – 54, rte. des acacias – CH-1227 Geneva, Switzerland
(6) LORIA-CNRS – Campus Scientifique - BP 239 – Vandoeuvre lès Nancy Cedex, France
(7) IRMC, Bardo Center, B.4, Appt. 25, Le BARDO – Tunis, Tunisie
www.speech.kth.se

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 (French-English) has been created, which includes texts of different genres, with various degrees of difficulty for the alignment task. Second, significant methodological progress was made both on the evaluation protocols and metrics, and the algorithms used by the different systems. In the second period, now underway, ARCADE has opened to a larger number of teams and to 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 (memory-based translation, extraction of multilingual lexical and terminological resources, etc.) (Brown et al., 1991; Gale and Church, 1991; 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. 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). In addition, most research efforts were directed towards the easiest problem, that of sentence-to-sentence alignment. Alignment at the word and term level, which is extremely useful for applications such as lexical resource extraction, is still a largely open research avenue.

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 such improvement can be boosted by systematic evaluation. However, before the ARCADE project started, there was no formal evaluation exercise 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 in order to promote research in the field of multilingual 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).

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 now, 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 texts was preferred to the collection of a very big amount of similar data.

3.1 Format

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

3.2 JOC

The JOC corpus is composed of records of questions and answers regarding European Community matters. The data is regularly published as one section of the C Series of the Official Journal of the European Community in all its official languages. This corpus, which was collected and prepared within the MLCC and MULTEXT projects, contains written questions asked by members of the European Parliament on a wide variety of topics and corresponding answers from the European Commission in 9 parallel versions. The total size of the corpus is approximately 10 million words (ca. 1.1 million words per language), and the texts date from the year 1993. 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 totalling close to 300 000 words per language; 2) **SCIENCE**, five scientific articles of about 50 0 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 translations are much freer than in the other types of texts. The English version is slightly abridged, which poses interesting problems 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. 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 at 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 finer-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 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 La phrase numéro un.	t_1 This is the first sentence.
s_2 La phrase numéro deux qui ressemble à la 1ère.	t_2 This is 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 *recall* of alignment A 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 *precision* of alignment A 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 total of those 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): $F = 2.(recall \times precision)/(recall + precision)$. Let us assume the following proposed alignment:

s_1 La phrase numéro un.	t_1 This is the first sentence.
	t_2 This is the 2nd sentence.
s_2 La 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: $A \cap A_r = \{(\{s_1\}, \{t_1\})\}$. Recall and precision of alignment A with respect to A_r are $1/2 = 0.50$ and $1/3 = 0.33$ respectively. The F-measure is 0.40.

Recall and precision as defined above are rather severe. 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 that $A = \{a_1, a_2, \ldots, a_m\}$ and $A_r = \{ar_1, ar_2, \ldots, ar_n\}$, with $a_i = (as_i, at_i)$ and $ar_j = (ars_j, art_j)$, we can derive the following sentence-to-sentence alignments: $A' = \bigcup_i (as_i \times at_i)$ and $A'_r = \bigcup_j (ars_j \times art_j)$ Sentence-level recall and precision can thus be defined in the following way: $recall = |A' \cap A'_r|/|A'_r|$ and $precision = |A' \cap A'_r|/|A'|$.

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

5 Systems tested

Six systems were tested (the RALI team presented two different systems).

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 system 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 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 the Jacal system, the LIA system uses a pre-processing step involving cognate recognition which restricts the search space but in a less strict way than Jacal. Then, sentence alignment is achieved through dynamic programming, using a score function which combines several kinds of information: 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 with the amount 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 close to that of Gale & Church.

6 Results

Four sets of recall/precision measures were computed on the alignments proposed by the six systems for the various types of texts 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. These results call for some comment.

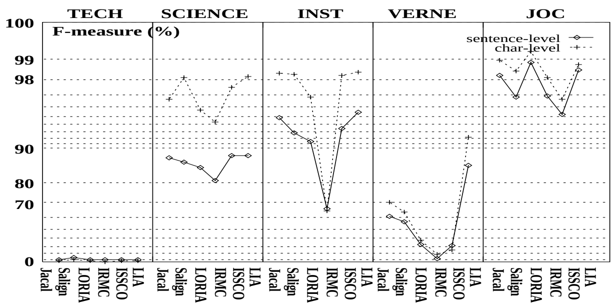

Figure 1: Global efficiency (average F-values) of the different systems (Jacal, Salign, LORIA, IRMC, ISSCO, LIA) , by type of text (logarithmic scale).

First, note than the *Char* measures are higher that the *Align* measures. This seems to confirm that systems tend to fail on shorter sentences. In addition, in the BAF corpus the reference alignment often combines several 1-1 alignments in a single n-n alignment, for practical reasons owing to the sentence segmentation process. This results in lowering the *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, which reflect a translation strategy based on speed and absolute fidelity. In addition, this corpus contains numerous data that are unmodified by the translation process (proper names, dates, etc.) and can be used as anchor points by some systems. Note that the LORIA system achieves a slightly better performance than others on this corpus, mainly because it is able to carry out a structure-alignment on this corpus, in which paragraph and divisions are explicitly marked.

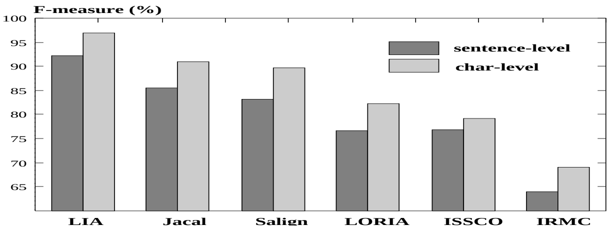

Figure 2: Final ranking on the systems (average F-values).

On the other hand, the VERNE corpus receives the worst results, as well as being the corpus on which the results are the most divergent across systems (22% to 90% char-precision). These poor results can be explained by the literary nature of the corpus, where translation is freer and more interpretative. In addition the English version is slightly abridged and the occasional missing sentences result in de-synchronization in most systems. Nevertheless, the LIA system still achieves a satisfactory level of performance (90% recall and 94% char-precision), which can be explained by the efficiency of the word-based pre-alignment step it uses, as well as the scoring function used to rank the candidate bisegments.

Significant discrepancy can also be 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, their definitions appear in different order in the two languages. This portion of text was not manually aligned in the reference, which results in an enormous bisegment (250-250) that dramatically 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 action was sentence alignment, it is probably not fair to take into account the TECH corpus in the final grading of systems.

The final ranking off all systems is given in Figure 2, in terms of the *Sent* and *Char* F-measures, and excluding the TECH corpus. The LIA system obtains the best average results, and shows good stability across texts, which is an 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 Elliott Macklovitch for his fruitful comments on this paper.

References

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.

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.

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.

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

W. A. Gale and K. 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.

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

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.

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