Evaluation of multilingual text alignment systems: the ARCADE II project

Yun-Chuang Chiao[1], Olivier Kraif[2], Dominique Laurent[3], Thi Minh Huyen Nguyen[4], Nasredine Semmar[5], François Stuck[6], Jean Véronis[7], Wajdi Zaghouani[8]

1. ELDA, 55-57, Rue Briallat Savarin, 75013 Paris France
2. LIDILEM, Université Stendhal Grenoble 3, 1180, Av. Centrale, 38400 Saint Martin d'Hères France
3. SYNAPSE Développement, 33, Rue Maynard, 31000 Toulouse France
4. LORIA, 615, Rue du Jardin Botanique, 54600 Villers-lès-Nancy France
5. LIST CEA de Fontenay aux Roses, 18, Route du Panorama, 92265 Fontenay aux Roses France
6. CRIM-INALCO, 2 Rue de Lille, 75007 Paris France
7. DELIC Université de Provence, 29, Av. Robert Schuman 13100 Aix-en-Provence France
8. European Commission, IPSC, Language technology group/JRC - T.P. 267 I - 21020 Ispra (VA) Italy

E-mail: chiao@elda.org, Olivier.Kraif@u-grenoble3.fr, dlaurent@synapse-fr.com, thi-minh-huyen.nguyen@loria.fr, nasredine.semmar@cea.fr, fstuck@inalco.fr, Jean.Veronis@up.univ-mrs.fr, wajdi.zaghouani@umontreal.ca

Abstract

This paper describes the ARCADE II project, concerned with the evaluation of parallel text alignment systems. The ARCADE II project aims at exploring the techniques of multilingual text alignment through a fine evaluation of the existing techniques and the development of new alignment methods. The evaluation campaign consists of two tracks devoted to the evaluation of alignment at sentence and word level respectively. It differs from ARCADE I in the multilingual aspect and the investigation of lexical alignment.

1. Introduction

With the rising importance of multilingualism in language industries, parallel corpora, consisting of source texts along with their translations into other languages, have become key resources for the development of natural language processing tools. The applications based upon parallel corpora are growing in number: multilingual lexicography and terminology, machine and human translation, cross-language information retrieval, etc.

The ARCADE I project (Véronis & Langlais, 2000), started in 1995 and ended in 1999, was designed to provide standard methods for the evaluation and comparison of parallel text alignment systems. The ARCADE evaluation exercise has allowed important methodological advances in the field for sentence alignment and a limited form for word alignment (“translation spotting”). The results include methods and tools for the generation of reference data and a set of measures for system performance assessment. In addition, a large standardized bilingual corpus has been constructed and can be used as a gold standard in future evaluation.

However the ARCADE I project had a few limitations. Only one language pair was tested (i.e., French-English) while it is well-known that other languages, especially languages with a non-Latin script (e.g., Arabic, Chinese, etc.) alignment remains an important issue to be addressed. French and English are relatively close to each other in terms of words, compounds and expressions. For other languages, such as Chinese, the notion of words and compounds is significantly different from that of European language.

The ARCADE II[1] project is one of the components of the EVALDA project in the Technolangue[2] framework financed by the French Research Ministry. The project aims at exploring the techniques of multilingual text alignment through a fine evaluation of the existing techniques and the development of new alignment methods. ARCADE II consists of two tracks devoted to the evaluation of alignment at sentence and word level respectively.

The ARCADE II differs from ARCADE I project in the multilingual aspect and the type of alignment addressed. Using French as the pivot language, 10 language pairs have been studied in ARCADE II: English, German, Italian and Spanish for western European languages; Arabic, Chinese, Greek, Japanese, Persian and Russian for more distant languages using non-Latin scripts. Sentence-level alignment was tested on each of these pairs, but, as opposed to the ARCADE I exercise, a subset of raw data (not pre-segmented in sentences) was also used. The word-level task took the form of a named-entity alignment for the Arabic-French pair.

2. Multilingual Parallel Corpora

One of the main results of ARCADE II has been to produce multilingual reference corpora in different languages. It is important to mention that until ARCADE II, there have been few projects of evaluation of parallel text alignment systems, notably the ARCADE I project and the Blinker project (Melamed, 1998), both deal with French-English alignment. Other works on word alignment evaluation were restricted to two or three languages pairs (Mihalcea and Pedersen, 2003; Martin et al., 2005). There were no formal evaluation exercises for multilingual parallel text alignment.

2.1. Data Format

Both western European languages and distant languages parts of the ARCADE II corpus (described below) are XML and UTF-8 encoded.

[1] http://www.up.univ-mrs.fr/veronis/arcade/index.html

[2] http://www.technolangue.net

2.2. JOC Corpus

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, written questions asked by members of the European Parliament on a variety of topics and the corresponding answers from the European Commission.

The part used in ARCADE II consists of the same subset of ca. 1 million words in English, French, German, Italian and Spanish (i.e. 5 million words altogether). English, German, Italian and Spanish were aligned to their French counterpart at the sentence and paragraph level. The corpus was converted to UTF-8 and XML format.

2.3. MD Corpus

The MD corpus consists of news articles from the French monthly newspaper Le Monde Diplomatique, which is translated and distributed in a number of languages around the world. The MD corpus contains 150 Arabic (Ar) texts aligned to French at the sentence level; about 50 aligned text pairs with French as pivot language for Russian (Ru), Chinese (Zh), Japanese (Ja), Greek (El) and Persian (Fa). A subset of French and Arabic parallel texts with named entity phrases hand-tagged was also provided for the word alignment task. Table 1 summarizes some statistics of the MD corpus.

	# doc.	# Kwords		# Kseg.		#Kalign.
Fr-Ar	150	517	403	14	11	11
Fr-Zh	59	197	-	5.2	5.5	4.45
Fr-El	50	179	190	4.3	4.4	4.37
Fr-Ja	52	240	-	5.7	6.1	5.51
Fr-Fa	53	214	220	5.2	5.3	4.61
Fr-Ru	50	173	158	4.2	4.2	4

Table1. Statistics on the MD corpus: number of documents, Kwords, Ksentences and aligned Ksentences.

3. Sentence Alignment task

Two subtasks have been defined, the systems being encouraged to participate in both:

- Segmented corpus. Sentence segmented multilingual texts were provided to the participants.
- Raw corpus. Multilingual texts were provided without sentence segmentation.

In both cases, the source and target texts were given to the participants, who had to return the aligned version within a fixed time span.

3.1. Evaluation metrics

As in ARCADE I, recall and precision were used to evaluate the quality of a given alignment with respect to a reference. The simplest way is to compute these measures by counting the number of correct alignments. However, some system and reference alignments can be partially correct, recall and precision as defined above are rather severe.

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

In ARCADE II, we decided to compute recall, precision and F-measure (Van Rijsbergen, 1979) only at the character level. Previous results in ARCADE I have shown that there is a strong correlation between the results obtained by using *word* and *character* granularities. It has also been noted that the *character* granularity, which is independent from sentence segmentation, seems the most convenient measure for the evaluation of alignment (Langlais et al., 1997).

3.2. Western European Language Alignment

Two subsets of (1) sentence-aligned and (2) raw texts were provided to participants for the following language pairs: English-French, German-French, Italian-French and Spanish-French.

3.2.1. Participants

Four systems were evaluated (P1, P2, P3, P6); one of them (P3), restricted to the French-English alignment, had format problems that prevented us from including the results in this paper.

3.2.2. Results

Recall, precision and F-measures were computed using *character* granularity for all western European language pairs, and for both sentenced-aligned and raw corpora. The overall results for all systems are given in Figure 1 in terms of average F-value for all language pairs.

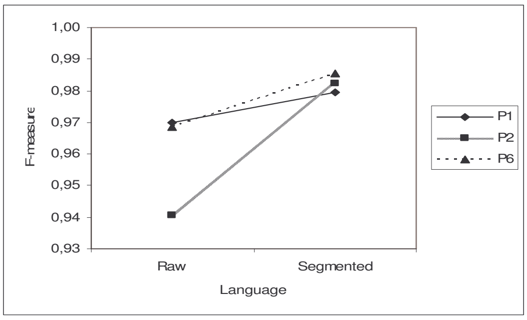

Figure1. F-measure for western European languages

The next figures show the split up per language for the segmented corpus (figure 2) and the raw corpus (figure 3).

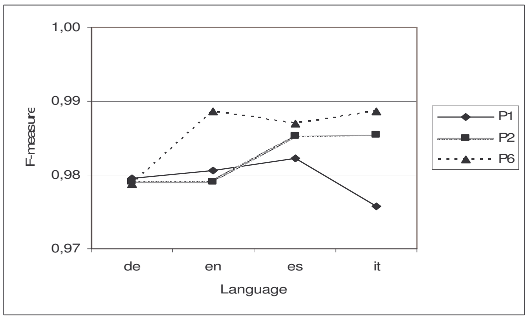

Figure2. Western European languages (segmented corpus)

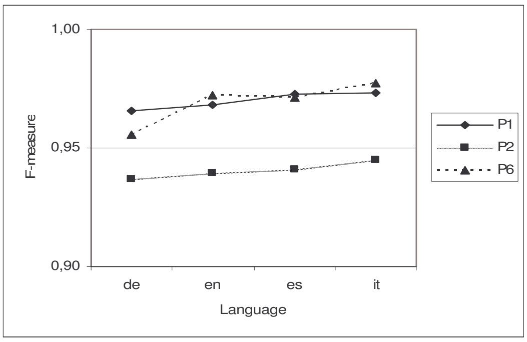

Figure3. Western European languages (raw corpus)

3.2.3. Discussion

The results show that the alignment task is much more difficult on raw corpora. On segmented corpora the average F-measure (for all languages) is around .98, whereas, on raw corpora it is around .97 for two systems, and goes as low as 0.94 for the system (P2). The split up per language shows that German is more difficult to align. On the segmented corpus, for example, the best system (P6) achieves results close to .99 in terms of F-measure for English, Italian and Spanish, but below .98 for German.

3.3. Distant Language Alignment

Sentence-aligned and raw parallel texts from the MD corpus[3] were provided to participants for the following language pairs: Arabic-French, Chinese-French, Greek-French, Japanese-French, Persian-French and Russian-French.

3.3.1. Participants

Two systems (P1, P2) were evaluated; the system P1 was restricted to the sentence-segmented text alignment.

3.3.2. Results

Recall, precision and F-measures were computed using *character* granularity for all distant language pairs, and for both sentenced-aligned and raw corpora. The evaluation results on the raw corpus for two systems are given in Figure 4 with the global efficiency (average F-values for all language pairs). Figure 5 shows the comparison of two systems for the segmented corpus.

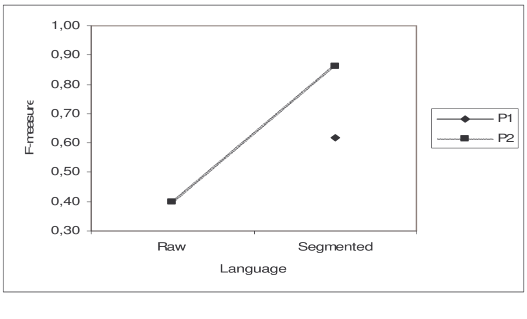

Figure4. F-measure for distant languages

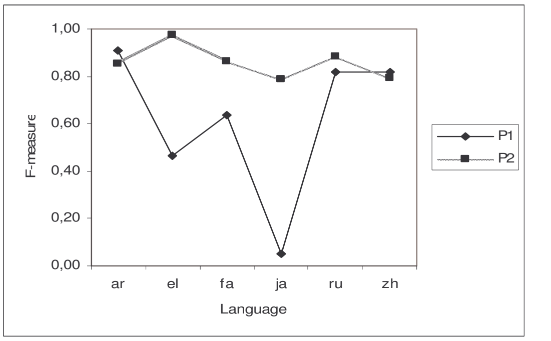

Figure5. Distant languages (segmented corpus)

3.3.3. Discussion

The results are much more modest than those concerning western European languages. The best results are achieved by P2 for Greek on the segmented corpus (F = 0.976), but the average for that system for all languages is only 0.871, thus much lower that the results obtained on western European languages. Sentence segmentation is obviously a very difficult task on distant languages using non-Latin scripts, since P1 was no able to perform it, and the results for P2 on the raw corpus are very modest (F = 0.421 on average).

4. Word Alignment task

Evaluation of words alignment between parallel texts presents more difficulties than the evaluation of sentence alignment (Véronis and Langlais, 2000), given the differences in word order between languages, the difference in part-of-speech and syntactic structure between the source and its translation, the discontinuity of multi-token expressions, etc. As a result, research is less advanced than in the area of sentence alignment.

It is not even entirely clear how some words, such as function words should be aligned when they do not have a direct counterpart in the other language. It was decided that the word alignment part of ARCADE II would be devoted to methodological issues and reference corpus construction; a small competition with a restricted set of participants was however conducted. The task proposed as a starting point in the ARCADE II was the identification of named entity phrases translation in parallel text.

French	Arabic
…Et ceux de **Forgeval** à **Valenciennes** (**Nord**), qui envisageaient d'incendier leur usine, ou encore ceux de **Bertrand Faure**, qui brisèrent des machines d'atelier à l'annonce de la fermeture de leur usine…	و الآخرون بتفجير براميل غاز. كما نذكر عمال **فورجيفال** في **فالانسيان** في **الشمال** الذين قرروا إحراق معملهم ، أو عمال **برتران فور**

Table2. Example of named-entity alignment

The identification of named entity translations can be seen as a sub-problem of full alignment, that of translation spotting. Given a particular word or expression in the source text, it consists in detecting its translation in the target text. An obvious application is the highlighting of translations for particular words on parallel texts presented

[3] Due to the nature of its content (news) and the diversity of the editorial sources of the translated versions, this corpus can be characterized by its inherent flaws (missing and merged segments, translation errors, etc.)

on a screen, as in multilingual concordance (Table 2), and the automatic construction of multilingual lexicons.

The scenario of this task was defined as follows: given a set of named entity tagged French texts, the participants had to identify the translation of the French named entities phrases in the untagged Arabic parallel texts.

4.1. Reference corpus

A subset of MD corpus was used in this experiment. For the purposes of this experiment, we hand annotated 60 French-Arabic parallel texts, designated as reference corpus. The named entities were tagged according the ESTER[4] named entity guidelines (Le Meur et al., 2004). A single annotator was used[5]. The French subset of the reference corpus is composed of 72,990 words and 3,639 named entity phrases were annotated. For the Arabic subset, 2,924 named entity phrases were annotated. Some statistics on the reference corpus are presented in Table 3.

	Pers.	Loc.	Org.	GSP.	Time
Fr.	546	44	397	1140	375
Ar.	550	44	368	1133	372

Table3. Number of major named entity types (person, location, organization, social-political grouping and time) tagged in the reference corpus.

Many difficulties remain for the constitution of a reference corpus and for the evaluation. In fact, a single-token word may have a multi-word unit as its counterpart in the target (one-to-many alignments), either for lexical reasons or grammatical ones; conversely, the source word can be part of an expression which is translated as a whole (many-to-one alignments). Different types of named entities are translated differently (Table 4). In such cases, the exact correspondence between the two texts is particularly difficult to be identified.

Named entity type	French	Arabic
Location	Pyongyang	كوريا الجنوبية
Temp. expression	1920	العشرينات

Table4. Examples of named entity phrases annotated in the reference corpus: *Pyongyang* was translated by *North Korea* and *1920* by *20's*.

4.2. Participants

Two systems were tested (P2 and P7), the output of the system P7 was restricted to few types of named entities (person and organization).

4.3. Evaluation Results

A simple algorithm was first applied to count the number of matches between the output of systems and the reference corpus. Initially, only exact matches were considered. Table 5 shows the preliminary results.

	P2	**P7**
# of NE identified	2,697	699
# of exact matches	82	98

Table5. Results of exact matches for systems P2 and P7.

The human evaluation is still underway at the time of writing this paper and will be available for the oral presentation.

4.4. Discussion

Given the lack of prior experience and the short time span, the scope of the named entity alignment task was limited. The exercise was however useful for defining and testing a protocol and metrics. Manual annotation of French-Arabic named entity alignment has been tested and a preliminary set of guidelines has been drafted.

5. Conclusions and Future Research

Although limited in too many respects and a short time-span, the ARCADE II exercise enabled the set of participants to build some expertise in handling new sets of languages, especially distant languages using non-Latin scripts. It also allowed the definition of an evaluation methodology and the production of reference resources which will be usable in future campaigns. It has shown that multilingual sentence-level alignment is a well-mastered task for languages based on the Latin script (although German poses some problems), and that the main difficulty in alignment in that of segmenting correctly the source text in sentences. Non-Latin scripts are still the source of many difficulties and results have been modest, both for sentence and named-entity alignment.

Acknowledgements

We would like to thank Olivier Hamon for his help and all the human evaluators.

6. References

Le Meur C., Galliano S. and Geoffrois E. (2004). ESTER Convention d'annotations en Entités Nommées.

Langlais P., Simard M. and Véronis J. (1997). ARCADE Methods and Practical Issues in Evaluating Alignment Techniques. Technical report. http://www.up.univ-mrs.fr/veronis/arcade/arcade1/report1-en/index.html

Martin J., Mihalcea R. and Pedersen T. (2005). Word Alignment for Languages with Scarce Resources. In Proceedings of the ACL 2005 Workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond", Ann Arbor, MI.

Mihalcea R. and Pedersen T. (2003). An Evaluation Exercise for Word Alignment. In Proceedings of the HLT-NAACL 2003 Workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond", Edmonton, Canada. pp. 1-10.

Rijsbergen CJ. Van. (1979): Information Retrieval. Butterworths, London (UK).

Véronis J. and Langlais P. (2000). Evaluation of parallel text alignment systems. The ARCADE project In N. Ide and J. Véronis (eds.): Parallel Text Processing: Alignment and Use of Translation corpora. Kluwer Academic Publishers. Chapter 19, pp. 369-388.

[4] French national campaign for automatic broadcast news transcriptions systems. (www.technolangue.net/article60.html)

[5] It may be rightfully argued that multiple annotators should be used.