Corpus Aligner (CorAl) Evaluation on English-Croatian Parallel Corpora

Sanja Seljan[1], Marko Tadić[2], Željko Agić[1]

Jan Šnajder[3], Bojana Dalbelo Bašić[3], Vjekoslav Osmann[3]

[1] Department of Information Sciences
[2] Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{sanja.seljan, zeljko.agic, marko.tadic}@ffzg.hr

[3]Knowledge Technologies Laboratory
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
{jan.snajder, bojana.dalbelo, vjekoslav.osmann}@fer.hr

Abstract

An increasing demand for new language resources of recent EU members and accessing countries has in turn initiated the development of different language tools and resources, such as alignment tools and corresponding translation memories for new languages pairs. The primary goal of this paper is to provide a description of a free sentence alignment tool CorAl (Corpus Aligner), developed at the Faculty of Electrical Engineering and Computing, University of Zagreb. The tool performs paragraph alignment at the first step of the alignment process, which is followed by sentence alignment. Description of the tool is followed by its evaluation. The paper describes an experiment with applying the CorAl aligner to a English-Croatian parallel corpus of legislative domain using metrics of precision, recall and F1-measure. Results are discussed and the concluding sections discuss future directions of CorAl development.

1. Introduction

Parallel corpora and particularly sentence-aligned multilingual corpora can be very effectively used as language resources in numerous research experiments, e.g. in creation of new sentence or word aligned resources, for computer-assisted translation, machine translation, multilingual information retrieval, language learning, multilingual terminology extraction and semantic networks (Seljan et al., 2007). The parallel corpus with the largest amount of tokens and languages covered so far – the JRC Acquis Communautaire (EC-DG-JRC, 2007), (Steinberger et al., 2006) – has already proven its value in experiments dealing with information retrieval, machine learning, statistical machine translation and in creation of translation memories and other electronic resources.

Croatia is an currently an candidate country for EU membership and the Croatian language appears more frequently in the international cooperation activities at the economic, cultural and political level in the EU. In such an environment, the use of translation tools becomes indispensable. In order to cope with expected overload in this direction, the need for shared tools and resources has become evident. For the Croatian language, which could be regarded as less spread language with only 5 million speakers, but having rich and valuable written communication with European countries, creation of multilingual parallel resources would add to the international communication, particularly translation, but also to the preservation of the national and European cultural identity.

In order to speed up the creation of aligned parallel corpora, a free alignment tool CorAl has been created at the University of Zagreb, Faculty of Electrical Engineering in the Knowledge Technology Laboratory.

CorAl is a tool for the alignment of parallel corpora. The starting idea was to create a tool for the alignment of parallel corpora which would cater to the specific needs of researchers, but also translators that are not motivated by research goals but simply want to develop their translation memories (TMs) in a straightforward manner. Available commercial solutions such as WinAlign (WinAlign, 2007) were not suited for the defined research tasks and were not available for free, so a new solution was developed from scratch. For research purposes there exist several, even on-line tools such as e.g. Uplug (Tiedemann, 1999) and also alignment visualisation tool (Tufiş et al. 2008) but they lack the user friendliness and generality of the graphical user interface (GUI). CorAl was inspired by existing and available tools, but it also added new functionality and provided new interface features that make usage or CorAl easy even for non-programmers.

2. The experiment

In this chapter, we provide a brief description of language resources and tools used in the experiment, along with the experiment framework.

2.1 CorAl aligner

CorAl was implemented entirely in Java, allowing portability to virtually any machine with Java Runtime Environment installed. The Model-View-Controller

(MVC) architectural pattern was used to isolate the user interface from background business logic. This separation implies limited dependence between the interface and the business model and provides that making changes to the interface is relatively simple. The graphical user interface was developed using Swing, a widget toolkit for Java.

In the case that the input texts are in the plain text format, thus carrying no sentence delimiters, CorAl's sentence segmentation module can be used for automatic sentence segmentation. For successful sentence segmentation, a list of abbreviations serving as exceptions in the appropriate language is required. Manual on-screen segmentation and merging of sentences is also supported.

At this level of development, CorAl offers two main modes of text alignment.

The first is a completely manual mode for creating sentence alignments. This mode of operation makes full use of the advanced GUI developed specifically for the ergonomic manual creation of sentence alignments.

The second is an automated mode which aligns, first the paragraphs, and then the sentences inside of each paragraph of the parallel text. The automated mode uses our own Java implementation of the well-know dynamic programming approach in the Gale-Church sentence alignment algorithm (Gale & Church, 1993).

Since CorAl is modularly composed, it is quite simple to add additional functionality such as phrase alignment and/or word alignment module, providing that applicable algorithms are available (e.g. Giza++ toolbox). This is certainly one of the desired directions of further improvement of the system.

2.2 Corpus

The corpus used for evaluation is the corpus of Croatian translation of the Acquis Communautaire documents (available at URL http://ccvista.taiex.be). The corpus consists of English and Croatian legislative documents, sentence-aligned using CorAl as the tool of choice for (semi-)manual alignment. The (semi-)manual alignment was done by a single person i.e. there was no inter-annotator agreement calculation needed. The (semi-)manually aligned corpus was thoroughly checked by another person and cleared of all sentence segmentation and alignment mistakes. This has provided us with the gold standard for evaluating the procedure of automatic sentence alignment.

The legislative English-Croatian subset consists out of 635 English-Croatian documents consisting of bylaws, regulations, and decisions. The documents consist of 1.8 Mw in English and corresponding 1.7 Mw in Croatian. Documents were provided in plain text format, manually aligned using CorAl aligner and used as a reference point when performing automatic alignment evaluation. Document stats are given in Table 1.

	Cro	Eng
Documents	635	635
Tokens	1,718,677	1,809,801

Table 1 Document stats

Legislative documents are included in the experiment because various statistical machine translation platforms, relying on sentence- and word-alignment preprocessing, are largely utilized exactly in tasks of translating legal documents, e.g. in the multilingual environment of the EU. It was therefore important to provide results of aligning English and Croatian in this domain in order to indicate the quality of the sentence alignment platform on which future research is about to build machine(-aided) translation systems.

2.3 Evaluation

As stated in the first section, among alignment tools implementing many standard and specialized algorithms, we chose to provide figures on English-Croatian pair using our own (semi-)automatic aligner CorAl. Beside previously explained reasons, there are two other reasons behind this choice: one is that it implements a standard Gale-Church algorithm that we wanted to evaluate on a language pair English-Croatian and the other is inclusion of this research into the joint research programme "Computational Linguistic Models and Language Technologies for Croatian" and its goals, described in detail by (Dalbelo Bašić et al., 2007). CorAl aligner is envisioned to be a default platform for sentence alignment (automatic and human assisted, i.e. semi-automatic) of language pairs Croatian-{any other language}. For that reason, evaluation was required in order to develop newer and better versions of the tool.

Evaluation method used in this experiment was highly influenced by the one presented in (Langlais et al., 1998). We have chosen the set of methods used during the ARCADE text alignment evaluation project as a starting point for our own experiment. Figure 1 provides an example on which evaluation techniques are illustrated.

A_R	s_1	*Ovo je prva rečenica.*	t_1	*This is the first sentence.*
	s_2	*Ovo je druga rečenica i nalik je prvoj*	t_2	*This is the second sentence.*
			t_3	*It looks like the first.*

A_S	s_1	*Ovo je prva rečenica.*	t_1	*This is the first sentence.*
			t_2	*This is the second sentence.*
	s_2	*Ovo je druga rečenica i nalik je prvoj.*	t_3	*It looks like the first.*

Figure 1 Example alignments

The formal description is as follows. We consider source text S and output text T as a sequence of alignments $\{s_1, \ldots, s_n\}$ and $\{t_1, \ldots, t_m\}$, respectively. An alignment A is then defined rather straightforward as a subset of Cartesian product of power-sets $2S \times 2T$. We then call the 3-tuple (S, T, A) a bi-text and each of its elements is called a bi-segment. Given these definitions, we set up two basic evaluation methods and consider two additional tweak or helper methods as proposed by (Langlais et al., 1998).

Recall and precision are easily defined on a bi-text as the number of correct assignments divided by the number of assignments in the reference alignment (recall) and in the system alignment (precision). Being that recall basically measures coverage alone and precision deals with counting correct alignments, F1-score (Rijsbergen, 1979) or harmonic mean is chosen for constraining these two outputs.

On example in Figure 2, the measures calculate as follows:

$$A_R = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_2, t_3\})\}$$

$$A_S = \{(\{s_1\}, \{t_1\}), (\{\}, \{t_2\}), (\{s_2\}, \{t_3\})\}$$
$$A_R \cap A_S = \{(\{s_1\}, \{t_1\})\}, |A_R \cap A_S| = 1; |A_R| = 2; |A_S| = 3$$
$$\text{Recall} = 0.50; \text{Precision} = 0.33; F_1\text{-measure} = 0.40$$

Being that this framework is rather harsh (an observer would intuitively state that the alignment is better than F1-measure indicates) and also rather high-level-oriented, we introduced, once again according to (Langlais et al., 1998) metrics, other and more finely-grained bi-segment subdivisions and cast the F1-measure framework onto them. In the presented example some segments are only partially correct, e.g. ($\{s_2\}$, $\{t_3\}$), which is the reason to measure recall and precision at the sentence level, and not at the alignment level.

Given alignments $A_R=\{ar_1,\ldots,ar_n\}$ and $A_S=\{a_1,\ldots,a_m\}$, with $a_i=(as_i,at_i)$ and $ar_j=(ars_j,art_j)$, sentence-to-sentence level metrics can also be defined.

Once again, on example set in Figure 2, the sets are defined and measures calculated as follows:

$$A'_R = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_2\}), (\{s_2\}, \{t_3\})\}$$
$$A'_S = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_3\})\}$$
$$A'_R \cap A'_S = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_3\})\}, |A'_R \cap A'_S| = 2;$$
$$|A'_R| = 3; |A'_S| = 2$$
$$\text{Recall} = 2/3 = 0.66; \text{Precision} = 2/2 = 1; F_1\text{-measure} = 0.80$$

It is now obvious that the sentence granularity and measure is more forgiving than the alignment granularity and that it is also somewhat closer to human intuition and evaluation. We thus chose sentence track F1-measure as a base for our experiment.

Method of (Langlais et al., 1998) suggests tuning the sentence track F1-measure by added granularity as A'_R and A'_S set cardinality could be expressed in terms of token count and character count. These tweaks are called word granularity and character granularity by (Langlais et al., 1998) and we chose to waive them for the purposes of this experiment. We find them somewhat useful, as they introduce reward to partial correctness of sentence alignment, but also judge them as inherent to the Gale-Church algorithm by default and therefore not to be of major effect to overall results. We proceed to results presentation for alignment track and sentence track evaluation in the following section.

3. Results and discussion

Evaluation results on alignment level and sentence level F1-measure track are provided in Table 2 for the legislative document corpus.

Track	Precision	Recall	F1-score
Alignment	0.977	0.971	0.974
Sentence	0.985	0.979	0.982

Table 2 Alignment accuracy

When considering results provided by (Gale & Church, 1993) for the core algorithm and results of (Langlais et al, 1998.) for various specific algorithms, implementing pre- and post-processing steps encapsulating Gale-Church algorithm, these results delivered by CorAl are rather predictable and expected with respect to the general properties of legislative texts. Being that Gale-Church algorithm is proven to work excellent in detecting one-on-one alignments and legislative texts provided in our test case are both quite small and straightforward in terms of manual alignment complexity, the figure of approximately 98.20 percent of correct alignments is not surprising. It should be noted that only a minor difference was observed between the alignment track and sentence track evaluation metric, once again due to the reduced complexity of the evaluation set. Namely, by definition of these two metrics, a more substantial difference should emerge when evaluating sentence alignment on texts with less one-to-one sentence alignments, because the sentence track metric would account for partial alignments in cases where the Gale-Church algorithm is known to introduce noise. However, this was not the case on English-Croatian legislative bi-texts.

4. Conclusions and future work

The work presented in this paper certainly leaves room for improvements. Results of this research would without doubt be better and more reliable with larger and/or annotated corpora, which then could be used for tasks at sub-sentence level, such as word alignment, terminology extraction, creation of thesauri, online dictionaries, semantic networks, etc. Also, different text genres are expected to show different results and this could also be one of directions for further research.

Integration of the Croatian language into this kind of multilingual resource, that JRC Acquis Parallel Corpus certainly is, by adding one more language should enable additional research on new cross-language relations.

Beyond the scope of building additional corpora and enriching the existing ones with additional linguistic annotation, technical improvements might include implementing pre- and post-processing steps around the core Gale-Church algorithm in order to handle possible non-one-to-one alignments with somewhat higher recall and precision. The Gale-Church algorithm itself, as a dynamic programming method, might enable additional tweaks or integration with other language pre-processing modules. Future research activities could also include alignment experiments at the lower linguistic level i.e. phrase and word level or they could include building basic language models and finally experimental systems for statistical machine translation.

5. Acknowledgements

The research within the project Accurat leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347. This work was also supported by the Croatian Ministry of Science, Education and Sports, under the grants 036-1300646-1986, 130-1300646-0909, 130-1300646-0645 and 130-1300646-1776.

6. References

Ceauşu Alexandru, Dan Ştefãnescu and Dan Tufiş. 2006. Acquis Communautaire Sentence Alignment using Support Vector Machines, Proceedings of the 5th

Language Resources and Evaluation Conference LREC2006, Genoa, Italy.

Dalbelo Bašić Bojana, Zdravko Dovedan, Ida Raffaelli, Sanja Seljan, Marko Tadić. 2007. Computational Linguistic Models and Language Technologies for Croatian. Proceedings of the 29th Information Technoloy Interfaces Conference. Cavtat, Croatia 521-528.

EC-DG-JRC (2007) The JRC-Acquis multilingual parallel corpus and Eurovoc (v. 3.0). JRC, Ispra, Italy. (availabe at http://wt.jrc.it/lt/Acquis/JRC-Acquis.3.0/doc/README_Acquis-Communautaire-corpus_JRC.html)

Gale William A. and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora, Computational Linguistics, 19(1):75–102.

Langlais Phillipe, Michel Simard, Jean Véronis. 1998. Methods and practical issues in evaluating alignment techniques. Proceedings of COLING-ACL98, Montreal, Canada 711-717.

Rijsbergen Cornelis. 1979. Information Retrieval, Butterworth-Heinemann, Newton, MA.

Seljan Sanja, Željko Agić and Marko Tadić. 2008. Evaluating Sentence Alignment on Croatian-English Parallel Corpora. Proceedings of the 6th International Conference Formal Approaches to South Slavic and Balkan Languages FASSBL. Dubrovnik, Croatia 101-108

Seljan Sanja, Angelina Gašpar and Damir Pavuna. 2007. Sentence Alignment as the Basis For Translation Memory Database. Proceedings of the INFuture2007 – The Future of Information Sciences: Digital Information and Heritage. Zagreb, Croatia 299-311.

Steinberger Ralf, Bruno Pouliquen, A. Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş and D. Varga. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. Proceedings of the 5th Language Resources and Evaluation Conference LREC2006.

Tadić Marko. 2000. Building the Croatian-English Parallel Corpus. Proceedings of the 2nd Language Resources and Evaluation Conference LREC2000, Athens, Greece 523-530.

Tiedeman, Jörg. 1999. Uplug - a modular corpus tool for parallel corpora. Parallel Corpora, Parallel Worlds. Proceedings of the Symposium on Parallel Corpora, Uppsala, Sweden.

WinAlign. 2007. Trados website listing applications (available at http://www.translationzone.com/en/products/sdltrados2007/applications/)