Sentence alignment in DPC: maximizing precision, minimizing human effort

Julia Trushkina[+], Lieve Macken*, Hans Paulussen[+]

K.U. Leuven – Campus Kortrijk, Belgium ([+])
Language and Translation Technology Team – Ghent, Belgium (*)
Yulia.Trushkina@kuleuven-kortrijk.be, Lieve.Macken@hogent.be, Hans.Paulussen@kuleuven-kortrijk.be

Abstract

A wide spectrum of multilingual applications have aligned parallel corpora as their prerequisite. The aim of the project described in this paper is to build a multilingual corpus where all sentences are aligned at very high precision with a minimal human effort involved. The experiments on a combination of sentence aligners with different underlying algorithms described in this paper showed that by verifying only those links which were not recognized by at least two aligners, an error rate can be reduced by 93.76% as compared to the performance of the best aligner. Such manual involvement concerned only a small portion of all data (6%). This significantly reduces a load of manual work necessary to achieve nearly 100% accuracy of alignment.

1. Introduction

A wide spectrum of multilingual applications have aligned parallel corpora as their prerequisite. These applications include, among others, machine translation (MT), especially corpus-based MT like statistical MT (Koehn 2005) and example-based MT (Carl & Way 2003), computer-assisted translation tools (Hutchins 2005), multilingual information extraction and computer-assisted language learning (Desmet & Paulussen 2005). More fundamental research in the fields of contrastive linguistics and translation studies (Baker 1996; Laviosa 2002; Olohan 2004) also profits from the use of parallel corpora.

For certain application (e.g. training machine translation systems) it is sufficient to extract only the 1:1 alignments (Moore 2002). Other applications however, require that all sentences in a corpus are aligned. These applications include, for example, translation studies and computer-assisted language learning.

A range of tools and algorithms is available for the task of sentence alignment, including, among others, sentence-length-based approaches (Gale and Church 1993), (Varga et al 2005), word-correspondence-based approaches (Melamed 1997), mixed approaches (Moore 2002). The performance of the tools varies for different types of texts and language pairs and normally, a manual verification step is necessary to guarantee high quality of the data.

The aim of the project described in this paper is to link all sentences of a corpus with very high precision but minimizing human effort. The paper describes experiments in which sentence alignment tools are combined. We present a formal evaluation of the tools and show that by combining outputs of aligners one can significantly reduce the amount of manual work necessary to achieve near 100% accuracy of alignment for the entire data set.

The article is organized as follows: the second section provides a short overview of the Dutch Parallel Corpus Project, in the framework of which the sentence alignment experiments have been carried out. The main part of the paper concentrates on the sentence alignment experiments: the tools used are presented and evaluated and a combined approach is described. Section 4 concludes the paper.

2. DPC Project

The aim of the Dutch Parallel Corpus project is to develop a high-quality annotated parallel corpus of ten million words for Dutch, French and English. At the moment of the abstract submission, the DPC project has just completed its second stage which concentrated on data alignment.

The DPC has the following features:

1. Balanced composition

   Since for different types of texts a different translation strategy is being adopted, the corpus is designed to represent as wide a range of written texts as possible. The text types include literary prose and non-fictional material, such as essayistic, journalistic, business, technical and policy texts. All text types will be equally distributed in representation of the corpus.

2. Quality control

   Three forms of quality control are envisaged for the DPC data: manual verification, spot-check, and automatic control procedures. This article provides details on how manual verification can be assisted by automatic control procedures on the sentence alignment task.

3. Sentence alignment
   The whole DPC corpus will be sentence aligned. A small part of the corpus will be additionally aligned on sub-sentence level.

4. Size
   The corpus will consist of ten million words.

5. Language pairs and translation directions
   The corpus consists of two bidirectional bilingual parts and one trilingual part (see Table 1).

EN	←	NL	→	FR
EN	↔	NL		
		NL	↔	FR

Table 1 DPC translation directions

6. Availability
   The corpus will be made available through the Dutch agency for Human Language Technology. Copyright clearance is being obtained for all samples included in the corpus.

A more detailed description of the project goals, applications and functionality can be found in (Macken et al 2007) and (Paulussen et al 2007).

3. Sentence Alignment within DPC

In sentence alignment, each sentence of the source language text is connected with the equivalent sentence or sentences of the target language text. The following alignment links are legitimate in the DPC project: *1:1, 1: many, many :1, many : many, 0 : 1, 1 : 0*. Zero alignments are created when no translation can be found for a sentence of either the source or the target language. Many-to-many alignments are legitimate in two cases: overlapping alignments and crossing alignments.

Tables 2 and 3 give examples of overlapping and crossing alignment cases. In both cases, multiple alignment 2:2 have to be created (S_1, S_2 vs. S'_1, S'_2).

Source language text	*Target language text*
S_1: A, B, C	S'_1: A′, B′
S_2: D, E	S'_2: C′, D′, E′

Table 2: An example of an overlapping alignment

Source language text	*Target language text*
S_1: A	S'_1: B′
S_2: B	S'_2: A′

Table 3: An example of a crossing alignment

A hybrid approach is used for sentence alignment of DPC data. The outputs of three aligners with different underlying heuristics are combined and then partially verified manually. The tools used in the experiments together with their evaluation are described below.

The **Vanilla aligner** (Danielsson and Ridings 1997) is an implementation of a sentence-length-based statistical approach of Gale and Church (1993). As input, the Vanilla aligner expects texts split into sentences and paragraphs. The numbers of paragraphs in source and target languages should be equal. The tool assumes that the paragraphs are aligned and finds sentence links within this paragraph alignment.

The **Smooth Injective Map Recognizer (SIMR)** developed by Melamed (1997) is a bitext mapping algorithm. By bitext, a text in two different languages is understood. The algorithm is based on word correspondences and relies on finding cognates (tokens with the same meaning and similar spelling) in a bitext to suggest word correspondences.

The **Microsoft Bilingual Aligner** developed by Moore (2002) uses a three-step hybrid approach to sentence alignment. In a first step, an initial alignment is established using the sentence-length-based approach. In the second step, sentences aligned in the previous stage with the highest probabilities serve as a basis for training a statistical word alignment model (Brown et al 1993). Finally, the corpus is realigned, augmenting the initial model with sentences aligned based on the word alignments. The aligner uses sentence-length and lexical correspondences, both of which are derived automatically. The aligner outputs only 1 : 1 links and disregards alignments which involve more than one sentence.

Performance of the three aligners have been evaluated against manually aligned data. Seven records of EUROPARL speeches in Dutch and English (1510 and 1316 sentences, respectively) have been used as a test set. The standard metrics of recall, precision and f-measure are defined as follows:

Precision = # correct alignments /
 # proposed alignments
Recall = # correct alignments /
 # reference alignments
F-measure = 2 * Recall * Precision /
 (Recall + Precision)

Table 4 summarizes the results of the evaluation.

	Recall	**Precision**	**F-measure**
Vanilla	95.96%	95.06%	95.51%
Microsoft	85.06%	94.83%	89.94%
SIMR	95.07%	92.98%	94.02%

Table 4. Evaluation of the DPC sentence aligners

The evaluation demonstrates the relative strengths of each aligner. Vanilla yields the highest results, but requires most manual involvement in the form of pre-processing paragraph alignment. The Microsoft aligner achieves a high precision on 1:1 alignments but neglects 1:many and many:1 alignments, which is harmful for this type of texts:

Europarl speeches contain rather long sentences and during translation the sentences are split into shorter ones. The SIMR aligner provides high accuracy with no manual pre-processing involved.

In order to further improve the alignment quality, a partial manual control is performed. In the output of the Vanilla aligner, all links which were not recognized by at least one other aligner, are marked. In our experiments, an average number of such links is 6% of the total test set. These non-shared links are checked, and, if necessary, corrected manually. No other links are changed.

The corrected output has been compared to a gold standard. The comparison has shown that manual control of 6% of the data resulted in 93.76% error rate reduction, yielding an accuracy of 99.72% (see Table 5).

	Recall	**Precision**	**F-measure**
Final output	99.68%	99.77%	99.72%

Table 5. Evaluation of the combined approach

An error analysis has shown that the remaining errors concern links which were recognized both by Vanilla and SIMR aligners and, therefore, were not marked to be checked manually. Below, typical errors of the three aligners are described.

Errors of the Vanilla aligner mainly concern links which contain more than two sentences in one language, for example 4:2, 3:1 or 4:1 alignments. Error analysis has shown that in this case, Vanilla prefers links with more equal lengths of sentences. Table 6 demonstrates examples of possible output of Vanilla for such cases.

Correct	**Vanilla**
4:2	2:1, 2:1
3:1	2:1, 1:0
4:1	1:0, 1:0, 2:1

Table 6. Examples of Vanilla errors

SIMR also makes this type of error, although less often. The most frequent type of error for SIMR is preference of zero alignments over 2:1 alignments.

As mentioned above, the main weak point of the Microsoft aligner is its neglect of 1:many and many:1 alignments.

4. Conclusion

The experiments on a combination of sentence aligners with different underlying algorithms showed that by verifying only those links that were not recognized by at least two aligners, an error rate can be reduced by 93.76% as compared to the performance of the best aligner. Such manual involvement concerned only a small portion of all data (6%). This significantly reduces a load of manual work necessary to achieve nearly 100% accuracy of alignment.

Our future plans include comparing different combinations of aligners on various text types and finding an optimal combination for each DPC text type. We will also compare results received on Dutch-English data to the performance of the tools on Dutch-French texts.

Acknowledgments

The DPC project is carried out within the STEVIN program, which is funded by the Dutch and Flemish Governments. In more personalized terms, DPC is also Piet Desmet, Maribel Montero Perez (KU Leuven – Campus Kortrijk), Lidia Rura and Willy Vandeweghe (School of Translation Studies, Hogeschool Ghent).

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead, in H. Somers (ed.), *Terminology, LSP and Translation*, pp. 175—186. Amsterdam, Philadelphia: Benjamins.

Brown, P. F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2), pp. 263—311

Carl, M. & A. Way (2003). Recent Advances in Example-Based Machine Translation. Dordrecht: Kluwer Academic Publishers.

Danielsson P. and D. Ridings (1997). Practical presentation of a vanilla aligner. Technical report, Sprakbanken, Institutionen for svenska spraket, Goteborgs universitet.

Desmet, P. & H. Paulussen (2005). CorpusCALL: opportunities and challenges, in *Proceedings of the CALICO congress*, Michigan State University, USA.

Gale W. A and K. W. Church (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1), pp.75—102.

Hutchins, J. (2005). Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation*, 17(1-2), pp. 5—38.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation, in *Proceedings of the Tenth Machine Translation Summit*, Phuket, Thailand, pp. 79—86.

Laviosa, S. (2002). Corpus-based Translation Studies. Theory, Findings, Applications. Amsterdam/New York: Rodopi.

Macken, L., J.Trushkina, and L.Rura. (2007). Dutch Parallel Corpus: MT Corpus and Translator's aid. In

Proceedings of the Machine Translation Summit XI, Copenhagen, Denmark, pp. 313—320.

Melamed, I.D. (1997). A Portable Algorithm for Mapping Bitext Correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, pp. 305—312.

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Machine Translation: From Research to Real Users* (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, pp. 135—244.

Olohan, M. (2004). Introducing Corpora in Translation Studies. London/New York: Routledge.

Paulussen, H., L. Macken, J. Trushkina, P. Desmet, and W. Vandeweghe (2006). Dutch Parallel Corpus: a multifunctional and multilingual corpus. *Cahiers de l'Institut de Linguistique de Louvain*, CILL, Louvain-La-Neuve, 32(1-4), pp. 269—285.

Varga Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh & Viktor Trón (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*. Borovets, Bulgaria, pp. 590—596.