Experiments on Processing Overlapping Parallel Corpora

Mark Fishel, Heiki-Jaan Kaalep

University of Tartu
J. Liivi 2, Tartu 50409
Estonia
fishel@ut.ee, hkaalep@ut.ee

Abstract

The number and sizes of parallel corpora keep growing, which makes it necessary to have automatic methods of processing them: combining, checking and improving corpora quality, etc. We here introduce a method which enables performing many of these by exploiting overlapping parallel corpora. The method finds the correspondence between sentence pairs in two corpora: first the corresponding language parts of the corpora are aligned and then the two resulting alignments are compared. The method takes into consideration slight differences in the source documents, different levels of segmentation of the input corpora, encoding differences and other aspects of the task. The paper describes two experiments conducted to test the method. In the first experiment, the Estonian-English part of the JRC-Acquis corpus was combined with another corpus of legislation texts. In the second experiment alternatively aligned versions of the JRC-Acquis are compared to each other with the example of all language pairs between English, Estonian and Latvian. Several additional conclusions about the corpora can be drawn from the results. The method proves to be effective for several parallel corpora processing tasks.

1. Introduction

The number and sizes of available parallel corpora keep growing – e.g. the Europarl corpus (Koehn, 2005) has doubled and the JRC-Acquis corpus (Steinberger et al., 2006) – tripled during 2007; recently a multilingual parallel corpus of movie subtitles was announced as part of the OPUS corpus (Tiedemann and Nygaard, 2004), etc. This suggests increasing necessity for automatic methods of evaluating and combining the available corpora, as well as improving their quality. The aim of the work, described in this article, is to satisfy this necessity.

Sometimes the source documents of two independently created parallel corpora overlap. Such situations are additionally troublesome since there's often difference in source document versions, formats, encoding, etc. In addition different levels of alignment exclude the possibility of direct comparison of the sentences of the corpora.

On the other hand overlapping parts can be used to automatically detect alignment errors. In case one of the corpora is known to be more accurate, the other one can be proofed against it. Different levels of alignment can be synchronized, so that some units of both corpora get additionally segmented.

Here we present a method for processing parallel corpora containing overlapping parts, along with its implementation. Its main objective is to improve parallel corpora quality by detecting alignment errors and to avoid duplicate entries while combining overlapping corpora. We further describe a set of experiments on applying the method to different parts of JRC-Acquis.

2. Overlapping Parallel Corpora

The type of corpora that the introduced method is meant for is independently created parallel corpora that share common source documents – either fully or partially. For instance, the Estonian-English part of the Ispra JRC-Acquis corpus[1] and the parallel corpus of the University of Tartu[2] have 2 thousand common source articles (Kaalep and Veskis, 2007). Also the Hunglish corpus (Varga et al., 2005) contains both EU legislation texts (potentially overlapping with JRC-Acquis) and movie subtitles (potentially overlapping with the OPUS corpus). We use the former example for one of our experiments, reserving the latter for future work.

Another set of experiments was conducted on JRC-Acquis itself, as it contains two alternative alignment versions: one done with the Vanilla[3] and another with the HunAlign aligner (Varga et al., 2005). In this case the overlapping is almost full – according to (Steinberger et al., 2006) in case the confidence threshold of the aligner was not met, the documents were excluded from the corpus. We used the three language pairs between English, Estonian and Latvian in the experiments. The selection was motivated by the difference of all three and also by the scarcity of resources and experiments on the latter two.

In case of UT and JRC-Acquis it is easy to determine the documents included in both corpora as these are augmented with CELEX codes. Nevertheless sentence comparison here is a non-trivial task due to several differences.

First of all, the source documents were retrieved at different times for both corpora, which means that files of JRC-Acquis contain several minor corrections. Also the way special characters (e.g. like in *õlu*, *liköör*, *šņabis*, ...) are encoded is different in the two corpora.

Next, the level of segmentation is different in both corpora: whereas UT is aligned on sentence level, JRC-Acquis is only segmented into paragraphs and these are aligned. Although according to (Steinberger et al., 2006) most of the paragraphs in the corpus consist of only one sentence, it still

[1] further referred to as JRC-Acquis

[2] `http://www.cl.ut.ee/korpused/paralleel/`, further referred to as UT

[3] `http://nl.ijs.si/telri/Vanilla/`

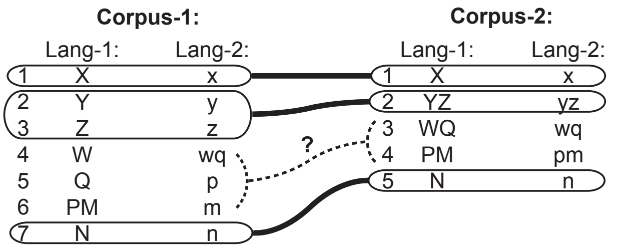

Figure 1: An example of correspondence between two parallel corpora chunks. The lines 4 to 6 of corpus-1 correspond to lines 3 to 4 of corpus-2, but contain erroneous alignments. Each letter stands for a phrase or a sentence. Solid lines indicate matches and dotted lines – mismatches

poses an additional problem for processing the corpora.

Also the two corpora were aligned with different methods. UT for example contains several shifts in the alignment; this type of error is more typical for Vanilla, and as a rule doesn't occur when using lexicalized aligners, such as HunAlign (Varga et al., 2005).

Finally, several text sections were left out when composing both corpora. Whereas in case of JRC-Acquis the missing parts can be extracted from the separately saved alignment, in case of UT this information is not provided. Therefore the easiest way of unifying the two corpora seems to be treating files of both as a linear input stream of sentences.

3. Method of Processing

The aim of the method introduced in this paper is to process two parallel corpora that have common source documents. Finding these common documents is treated as a separate task and is discussed, for instance, in (Kaalep and Veskis, 2007).

The method works by finding a correspondence between the sentence pairs of both parallel corpora; see figure 1 for an illustrative example. Having such a correspondence determined, it can be further used to combine the two parallel corpora in the preferred way (whereas repetitions in the resulting combination are avoided), to increase the segmentation level of one corpus on the account of the other, to check the accuracy of one corpus against the other, detect error locations for manually correcting them, etc.

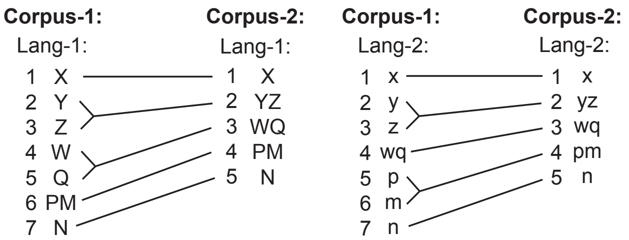

Figure 2: The first step towards finding a correspondence between two parallel corpora is to align the language parts separately

The following steps are taken to find such a correspondence. First the corresponding language parts are aligned separately with each other: in case of the first example of section 2 that would mean the Estonian parts of UT and JRC-Acquis aligned between themselves and the English parts – between themselves. This includes approximate sentence matching, in order to account for slight differences in the same sentence, coming from version, encoding or other differences. After the two alignments are found, these are compared to reveal mismatches between them. Finally the desired action is applied to the corpora using the comparison results: either a common corpus is generated, mismatch statistics are presented, and so on.

Consider the following example. Having the corpora from figure 1, first the lang-1 parts of corpus-1 and corpus-2 are aligned with each other and then the lang-2 parts (figure 2). Here several units of one side can match several on the other side. The alignments themselves are then compared using the same alignment techniques (figure 3), whereas now only 1-to-1 alignments are allowed. As a result, we obtain the correspondence of the sentence pairs of the two corpora, as in figure 1.

The main steps are explained in more detail in the following subsections.

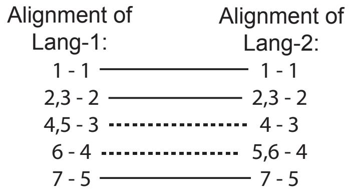

Figure 3: To find the actual correspondence of two parallel corpora the alignments of the two language parts are compared. In this example the 3rd and 4th lines do not match

3.1. Alignment of the Corresponding Language Parts

The first step is in essence very similar to the original task of bilingual sentence alignment itself. However, whereas the latter means comparing different languages and is therefore computationally difficult, in this case the task is much simpler, since both parts are in the same language and it suffices to compare the sentences directly by characters. The only problem is that instead of strict comparison of the sentences, here approximate comparison is required due to possible slight differences in different corpora.

For example, the sentence *x* in lang-2 of corpus-1 on figure 1 can have a typing error: "this is a shord sentence", and the same *x* in corpus-2 have the error corrected – "this is a short sentence". Although the two are obviously one and the same sentence, strict comparison would yield a mismatch between them.

The aligning task is therefore analogous to the longest common subsequence problem, where corpora units (e.g. sentences) are the elements and are matched approximately. In our implementation the alignment of the two texts is computed in an optimal way using edit distance. The cost of substituting a unit for another equals the distance between the two (which is explained in the next subsection) and the cost of insertion/deletion is always 1.

In addition to 1-to-1 substitution all N-to-M pairs are also considered up to a predefined limit (in our implementation – 10 by default). This enables detecting matching units even if the segmentation level is very different in the two corpora (e.g. matching together paragraphs and sentences).

3.2. Approximate Sentence Matching

(Kaalep and Veskis, 2007) use Levenshtein distance and check whether the distance between two sentences doesn't exceed 1% of the average length of the two sentences. Other string similarity metrics applied to written text include several from the edit distance family (the Needleman-Wunsch metric, the Smith-Waterman metric, etc), the Jaro metric and others.

In the current work we adopt the method from (Kaalep and Veskis, 2007), but use generalized edit distance instead of Levenshtein distance. For instance the weight of replacing/inserting digits is extremely high, so that e.g. sentences "article 3" and "article 5" will not be considered to match with no matter what edit distance percentage threshold. On the other hand operations on empty symbols (spaces, tabs) and punctuation have low weights. This allows to set the percentage threshold higher without adding obvious matching errors. Also character case is ignored during comparison.

3.3. Comparison of Separate Language Alignments

As soon as the language part alignments are obtained, their correspondence is to be determined. Although different language parts are to be compared here, only the alignments between unit numbers are compared, which again enables using direct comparison. In this case is accomplished again using edit distance, but this time with the simple Levenshtein distance of the alignment cells. Thus equality of the alignment elements indicates matching alignments while 1-to-1 inequality or 1-to-0/0-to-1 matches indicate mismatching alignments.

It is important to note that a mismatch between two alignments doesn't indicate, which of the corpora has an erroneous alignment; instead, it shows a potential spot, where at least one of the corpora has an error. In order to be used in automatic error correction, this setup requires that one of the corpora is preliminarily assumed to be more accurate. Alternatively, the spots can be manually post-processed, thus seeing which of the corpus contains the error and correcting it.

On the other hand a match between alignments also merely indicates that the two corpora have matching alignments. This can occur both in case of correct alignments and coinciding erroneous alignments, though the latter is less likely (depending on the used alignment method).

3.4. Implementation

After the sentence pairs get aligned it is still necessary to define the policy for sentence inclusion/exclusion from the resulting combined corpus. In the current implementation it is controlled by the user. It is possible to configure separately, whether to include sentences present in only one of the corpora, and the ones that match in both corpora.

In case of errors it is possible to include sentences from either one or the other corpora – in case one of the corpora is preliminarily known to contain less alignment errors. Alternatively it is possible to use one of the language-specific alignments or exclude the location of the error as a whole.

Logging of the alignments and error types is also supported. This enables error detection, and thus later human inspection and corpora post-processing.

The implementation is done in PERL and is available online[4].

4. Experiment-1: Combining Partially Overlapping Corpora

In this experiment we processed the overlapping parts of Hunaligned JRC-Acquis and UT; we used the older version (2.2) of JRC-Acquis as the newer (3.0) doesn't include the Hun-Align alignments yet, and the Vanilla alignment is much less accurate (Kaalep and Veskis, 2007).

The aim here was to obtain a common corpus with maximum size; therefore both the matching and the unique sentences were included from both input corpora. In order to include some sentences from alignment mismatches, it was necessary to decide which corpus was more accurate to use it as an error guideline. According to (Kaalep and Veskis, 2007) the potential error locations in the HunAlign version of JRC-Acquis are the 0 to N alignments, their predecessors and successors; all these were removed from the corpus and enhanced version was used.

In addition a second corpus was generated with only the matching sentences included and single and mismatched sentences left out. This would result in a smaller but much more accurate corpus – i.e. maximum accuracy in contrast to the maximum size of the first desired result.

Nr. of sentence pairs ($\cdot 10^3$)	**UT**	**JRC-Acquis**
Total:	92.3	67.6
Matched:	55.7	55.2
Single:	30.2	5.6
Mismatched alignments:	6.5	6.7
Max-size result:	98.2	98.2
Max-accuracy result:	55.7	55.7

Table 1: Output statistics of processing UT and JRC-Acquis

Results

The two resulting corpora were made available online together with the implementation of the introduced method. The output statistics of the processing results are showed in table 1.

Excluding the potentially erroneous alignments from JRC-Acquis reduced the number of sentences to 93% of the original. However, after processing the enhanced JRC-Acquis size was 102% of the original (111% of the reduced corpus). The size of the enhanced UT was 103% of the original. The size of the overlapping part grew to 106% of the

[4] http://ats.cs.ut.ee/smt/paralign/

UT part and 145% of the JRC-Acquis part. In total the resulting combined corpus size was 193% of UT and 161% of JRC-Acquis.

Based on the results in table 1, 60% of the UT sentences match with 82% of the JRC-Acquis sentences. The size of the maximum accuracy corpus is only slightly larger for JRC-Acquis than the matched sentences counted separately, which means that in the majority of cases the segmentation was deeper in UT.

It is theoretically possible that the matched sentence pairs include erroneous alignments; however in the current experiment a small randomly selected portion of the output was manually checked and no errors were discovered.

Match type	Nr. of occurrences
0-1	5621
1-0	30186
1-1	54723
1-2	59
2-1	426
2-2	1
3-1	5

Table 3: Types of sentence pair matches between UT and JRC-Acquis

Table 3 summarizes the types of matched alignments in the results (an N-M type means N sentence pairs in UT corresponding to M sentence in JRC-Acquis). These confirm both that the segmentation level in UT is slightly deeper (since there's more N-1 alignments than the other way around) and that the paragraphs of JRC-Acquis often contain only 1 sentence (since 1-1 alignments dominate).

5. Experiment-2: Comparing Different Alignments of the Same Corpus

In the second set of experiments the introduced method was applied to different alignments of the same parts of JRC-Acquis. The processed parts included three language pairs: English-Estonian, English-Latvian and Estonian-Latvian (unless otherwise specified, we further refer to these in the given order). The aim was to compare the different alignments and try to get a notion of the corpus accuracy; therefore no common corpora was generated.

Results

The results are displayed in table 2. The Estonian-Latvian part has a much higher percentage of matching sentences than the other two parts: 98% in both HunAlign and Vanilla versions versus 83% in the HunAlign and 86% – in the Vanilla version. It is possible that the Estonian-Latvian part contains much more coinciding errors, which would also cause the matching part to be larger. However a more desired explanation would be that this part is aligned more accurately.

In order to make sure we performed manual proofing of the results by randomly picking some files and checking whether the matching sentences reside in correct alignments and that mismatching sentences really include an alignment error[5].

Match type	Nr. of occurrences		
	En-Et	En-Lv	Et-Lv
0-1	3061	3076	661
1-0	1798	2005	158
1-1	251608	254743	315603
1-2	1	8	10
2-1	94	80	151

Table 4: Types of sentence pair matches between HunAlign and Vanilla versions of JRC-Acquis

None of the manually checked files contained coinciding errors in the Estonian-Latvian parts; in the other two parts mostly some two Estonian or Latvian sentences were erroneously grouped into one. An extract from the corpora (parts of documents with the CELEX number 31965R0079) along with the program output is displayed in figure 4

Table 4 summarizes the types of matching sentence pair alignments in all three experiments. Expectedly, most of the alignments are one-to-one, with rare two-to-one instances.

6. Conclusions and Future Work

We presented a method of automatic processing of overlapping parallel corpora. The method enables comparing corpora and finding mismatches in alignments, improving corpora quality both automatically and manually via post-processing and combining the input into a common corpus without including duplicate entries. The method is insensitive to minor differences in the aligned sentences, or to large sections missing from one of the corpora. It also takes into consideration possible differences in the level of segmentation.

A set of experiments of applying the method to the JRC-Acquis corpus was described. In the first experiment the Estonian-English part was combined with the parallel corpus of the University of Tartu. The results show that the latter has a higher level of segmentation but sometimes slightly lower alignment accuracy. Two common corpora were generated, based on the two: one with the maximum-size criterion (193% of the UT corpus and 161% of JRC-Acquis) and another with the maximum-accuracy criterium (60% and 80% of the overlapping parts of the UT and the JRC-Acquis corpora, respectively).

In the rest of the experiments the method was applied to the two alternative alignment versions of the JRC-Acquis: the HunAlign and the Vanilla version. Language pairs between three languages were tested: English, Estonian and Latvian. The results show that the Estonian-Latvian part of the corpus has a much higher number of matching sentence pairs (98% of both versions), which indicates good alignment quality.

Future work has several possibilities. Since the experiments were applied to the older version of JRC-Acquis, it would

[5] Special thanks to Zane Fishele for proofing the Estonian-Latvian and English-Latvian parts

Nr. of sentence pairs ($\cdot 10^3$)	**English-Estonian**		**English-Latvian**		**Estonian-Latvian**	
	HunAlign	**Vanilla**	**HunAlign**	**Vanilla**	**HunAlign**	**Vanilla**
Total:	301.6	295.2	304.0	295.7	322.4	321.6
Matched:	251.8	251.7	254.9	254.8	315.9	315.7
Single:	1.8	3.1	2.0	3.1	0.2	0.6
Mismatched alignments:	48.1	40.4	47.1	37.8	6.4	5.2

Table 2: Output statistics of processing UT and JRC-Acquis

Nr.	**HunAlign**			**Vanilla**		
	English	Estonian	Latvian	English	Estonian	Latvian
1	CHAPTER 1 Creation of a farm accountancy data network for the European Economic Community	ON VASTU VÕTNUD KÄESOLEVA MÄÄRUSE:	IR PIEŅĒMUSI ŠO REGULU.	CHAPTER 1 Creation of a farm accountancy data network for the European Economic Community	I PEATÜKK	I NODAĻA
2	Article 1	I PEATÜKK Euroopa Majandusühenduse põllumajandusliku raamatupidamise andmevõrgu loomine	I NODAĻA Eiropas Ekonomiskās kopienas lauku saimniecību grāmatvedības datu tīkla izveidošana	Article 1	Euroopa Majandusühenduse põllumajandusliku raamatupidamise andmevõrgu loomine	Eiropas Ekonomiskās kopienas lauku saimniecību grāmatvedības datu tīkla izveidošana
3	1. To meet the needs of the common agricultural policy, there . . .	Artikkel 1	1. pants	1. To meet the needs of the common agricultural policy, there . . .	Artikkel 1	1. pants

Figure 4: Extract from JRC-Acquis with all three languages and two alignment versions. It can be clearly seen even without knowing the used languages that there is an almost direct correspondence between Estonian and Latvian texts. The first and second pairs of Estonian-Latvian sentences in the Vanilla part match the second pair of sentences in the HunAlign part. On the other hand both the Estonian and the Latvian part form an analogical mismatch with the English part. In this case both the HunAlign and the Vanilla versions of English-Estonian and English-Latvian parts contain an alignment errors, however different ones

be interesting to process the newer and larger version of the corpus; this however requires the new HunAlign version to be released. Also the OPUS and Hunglish corpora can be experimented with.

Also the results of the first experiment can be used to manually post-process the corpus to correct the erroneous alignments.

Finding the corpus parts with common source documents is an open issue in the general case.

7. References

H.-J. Kaalep and K. Veskis. 2007. Comparing parallel corpora and evaluating their quality. In *Proceedings of MT Summit XI*, pages 275–279, Copenhagen, Denmark.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, Phuket, Thailand.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC'2006*, pages 2142–2147, Genoa, Italy.

J. Tiedemann and L. Nygaard. 2004. The opus corpus – parallel & free. In *Proceedings of LREC'2004*, pages 1183–1186, Lisbon, Portugal.

D. Varga, P. Halákcsy, A. Kornai, V. Nagy, L. Németh, and V. Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP-05*, pages 590–596, Borovets, Bulgaria.