Comparing Parallel Corpora and Evaluating their Quality

Heiki-Jaan Kaalep, Kaarel Veskis

Department of General Linguistics
University of Tartu
Liivi 2, 50409 Tartu
Estonia
{Heiki-Jaan.Kaalep, Kaarel.Veskis}@ut.ee

Abstract

The availability of partially overlapping parallel corpora for a language pair opens up opportunities for automatically comparing, evaluating and improving them. We compare and evaluate the alignment quality of two English-Estonian parallel corpora that have been created independently, but contain overlapping texts. We describe how to determine the overlapping parts and find their alignment similarities that allow us to economize on manual evaluation effort. We also suggest a feature that could be used instead of comparing and manual checking to predict the alignment correctness.

Introduction

Machine translation (MT) needs large parallel corpora. Their creation is a costly and labour-intensive process and there are never too many. It is no wonder that once a language pair has more than one parallel corpus, it is tempting to collate them in order to get a new and better corpus for MT.

It may happen that the original corpora overlap. This can be seen as a disadvantage because it makes collating more difficult: one should not include both of the overlapping parts, while recognising the overlappings is difficult because of the variety of conventions and formats of real corpora. However, overlapping also enables one to compare the corpora and thus evaluate them (half) automatically.

The latter is of considerable importance in view of the role of parallel corpora in MT, especially in evaluating the quality of the output.

When creating a new corpus, one often utilises methods and tools from previous work, evaluated on some corpus. However, the building material for the new corpus is different, and one should not rely on the assumption that the methods and tools yield exactly the same results (Langlais et al, 1998).

In the current paper we compare and evaluate two English-Estonian parallel corpora that have been created independently, but contain overlapping texts.

Aim of the Work

The quality of a parallel corpus is determined by the correctness of the parallel units, i.e. text snippets of the source language, aligned with their translations. If a parallel corpus is created automatically, then a 100% quality in this respect is nearly impossible to achieve. Worse still, it is extremely difficult to evaluate the correctness of alignments: we are facing the need to compare the respective meanings. In case of large parallel corpora, the question of evaluation is often not raised at all; see, for example, OPUS (Tiedemann, 2004), Europarl (Koehn, 2002), Czech–English (Bojar & Žabokrtský, 2006), Hungarian–English (Varga et al, 2005). Similarly, the LDC catalogue (http://www.ldc.upenn.edu/Catalog/) contains no information about the correctness of any of the parallel corpora it lists. In many corpora, it is the warning in the corpus documentation that alignments have been created automatically, or that the alignments may contain errors that indicates the estimation of the alignment quality. The only way to evaluate the alignments is to compare them manually on a smaller subset, but this is extremely labour-intensive (Samy et al, 2006; Singh & Husain, 2005).

We set ourselves two goals. The first goal was to evaluate the quality of the existing corpora, taking advantage of the existence of independently created alignments of the same source documents. Evaluating alignment quality of a corpus is different from evaluating the output of an aligning method: for a method, both precision and recall are important, while for a corpus, the only important characteristic is precision (instead of recall, we have corpus size).

The second goal was to find features predicting the quality of alignments in a document, so that we could estimate it even in the absence of a directly comparable, alternative aligned version from the other corpus.

Corpora

We had two English-Estonian corpora of legislation texts at our disposal.

UT corpus

The corpus of the University of Tartu at http://www.cl.ut.ee/korpused/paralleel/index.php?lang=en has two parts that represent the different source languages. One part is the Estonian legislation and its translations that make up 150,000 parallel units (sentences or list elements) in 400 texts, totaling 1.7 million tokens in Estonian and 2.9 million in English. The other part consists of EU legislative texts that make up 280,000 parallel units (sentences or list elements) in 4,000 texts, totaling 3.3 million tokens in Estonian and 4.9 million in English.

JRC-Acquis corpus

The corpus of EU legislation, *Acquis Communautaire* at http://langtech.jrc.it/JRC-Acquis.html is a parallel corpus of 21 European languages with an average of 8.8 million tokens in 7,600 texts per language. The Estonian part is 7.2 million and the English part 9.9 million tokens

(Steinberger et al, 2006). Having downloaded both the texts and scripts for producing the alignments from the address above and run the scripts, we found that the resulting English-Estonian sub-corpus contains 300,000 parallel units in 7,900 texts, totaling 4.6 million tokens in Estonian and 6.8 million in English. Thus this sub-corpus has considerably less tokens per language than the original corpus where all the language pairs are present.

Alignment

The alignments of both corpora were produced with language independent methods.

The UT corpus was aligned with Vanilla aligner (http://nl.ijs.si/telri/Vanilla/) that uses the algorithm from (Gale and Church, 1993). The strategy followed was similar to the one in creating the Europarl corpus (Koehn 2002): if the formal structures of the aligned units were too different, then these units were discarded from the resulting corpus altogether.

The aligning procedure passed through 3 stages: first, aligning chapters, parts and appendixes, then paragraphs, and finally sentences. The first stage was necessary because although it is common practice to preserve the formal structure of the original when translating legislative texts, one should not assume that the electronic versions of the original and the translation follow the same conventions for structural mark-up. One cannot be sure also that the original and translation both contain the same structural elements, e.g. date, heading, appendixes or tables, and in the same order.

After every stage, the numbering of parts of the documents or lists as the simplest anchor points was used for checking the alignments. If two parallel texts contained a different number of sections, articles or list items, or if the numbered elements were not parallel, then these texts were not included in the parallel corpus. It was assumed that in such cases the formal structure of the texts was too different from each other and that the simple method used would not yield trustworthy results.

Similarly, if the number of items (parts, paragraphs, sentences) to be aligned at a certain stage was too different in the parallel texts – one contained more than twice the number of the other – then the respective block was left aside completely.

The UT corpus contains only 1-1, 1-2, and 2-1 alignments. All the other alignments have been deleted.

The aligning of JRC-Acquis corpus proceeds in one stage – aligning the paragraphs. One can use two aligners, provided together with the aligning scripts from http://langtech.jrc.it/JRC-Acquis.html: Vanilla and HunAlign (http://mokk.bme.hu/resources/hunalign). HunAlign runs in three phases. First, it builds alignments using a simple similarity measure based on sentence lengths and the ratio of identical words. Number tokens are treated specially: similarity of the sets of number tokens in the two sentences is considered. The one-to-one segments found in this first round of alignment are fed to the second phase of the algorithm: a simple automatic lexicon-building. In the third phase the alignment is re-run, this time also considering similarity information based on the automatically constructed bilingual lexicon. (Steinberger et al, 2006).

The documentation at http://langtech.jrc.it/JRC-Acquis.html indicates that no part of the corpus has been omitted because of aligning difficulties. The corpus contains 0-alignments, i.e. parallel units where one of the translation pairs is empty.

The alignment level of the two corpora is different: paragraphs in JRC-Acquis and sentences in UT corpus. Therefore we had to use the version of UT corpus which had passed only through the first two stages, thus containing paragraph-level alignments.

Comparing and Evaluating

Overlapping Parts

We have to assume that both corpora contain texts, missing from the other. Thus the first task is to identify the overlapping part, as this is what enables us to automate the evaluation procedure and gives us information about the impact of the different alignment methods.

Identifying overlapping texts is not a trivial task, because deciding whether two texts are really the same is difficult. Texts with the same contents may have different lay-out, some parts of one text may be missing from the other or be in a different location. In case of legislative texts, for example, a date may be the first element of a text, or follow the heading, or precede the signature in the end, or be the last element of the text. Alternatively, texts that look very similar may actually be different. For example, a legislative act may repeat a previous one almost word by word. So the comparison of texts, even if we use approximate matching, is bound to give some incorrect results.

Fortunately, the EU documents have CELEX codes, i.e. codes used for identifying them. A translation has the same CELEX code as the original. It was possible to transform both the JRC-Acquis and the UT corpus so that an aligned text is a separate file with its CELEX code. This way it was possible to identify 2000 files in both corpora that should have the same content, and use these sub-corpora for evaluation.

Alignment Similarity (AS)

The first step was to find out how similar the parallel units of the corresponding files of different corpora are. Units that are similar are probably correct, while at least one of the dissimilar units is not. Manual checking of the alignments would be guided by the results of the first step.

We compared pair-wise the three aligned versions of the 2000 texts: the Vanilla and HunAlign versions of JRC-Acquis (“JRC Vanilla” and “JRC HunAlign”) and the UT corpus. We followed the same procedure in all the comparisons.

After the mark-up, accented letters and space characters had been unified across the JRC-Acquis and UT corpus, their texts still contained annoying small formal differences, e.g. non-standard accented letters, various ways of using brackets and punctuation marks, representing numbers and capitalization etc. In addition, sometimes a file contained more text than its counter-part in the other corpus. This could be caused by the differences in the sources from where the texts had been taken into the corpora, and it could be also the result of the aligning method used in the UT corpus which excluded the blocks that were difficult to align.

The corpora had also some differences in the Estonian wording, indicating that the texts represented different translations or different stages of editing.

All these small differences were considered as noise (the goal being to evaluate alignments, not translations).

In order to ignore the noise, we allowed the matching units to differ by a Levenshtein (edit) distance larger than zero.

The algorithm for finding alignment similarity (AS) of two texts was the following.

Step 1. Compare the texts without the alignment information to determine the size of their overlapping part. We chose to compare the English sides of the aligned files to find the number of similar English paragraphs (EnSim). It turned out that no English text had a completely similar counterpart in the other corpus; the similarity ranged from 0 to 99%.

The texts were first compared with the UNIX command *diff* with options that output both types of paragraphs: the perfectly overlapping ones and the dissimilar ones. The latter were compared once again, using Levenshtein distance in order to eliminate the impact of noise, and possibly added to the set of overlapping ones.

Step 2. Compare the aligned units (English and Estonian paragraphs concatenated) of the corresponding files to find size of the overlapping part, i.e. the number of similar aligned paragraphs (EnEtSim).

When deciding about the similarity of paragraphs, we allowed 2% of their characters to be dissimilar in step1, and 1% in step 2. This is because if we set the allowed dissimilarity to be the same in both cases, it may happen that two English paragraphs are less similar than allowed, but when concatenated with Estonian, the dissimilarity of the units is within allowed limits, and we have a counter-intuitive situation with EnEtSim > EnSim, i.e. that the number of correct alignments is bigger than the number of overlapping English paragraphs.

Step 3. Find the alignment similarity as the ratio between the number of similar parallel units (EnEtSim) and the number of similar monolingual units (EnSim) of two texts: AS = EnEtSim / EnSim.

AS can be used as an indicator of texts that deserve closer inspection.

Comparison of the JRC Vanilla and UT Corpus

We divided the 2000 texts into 8 groups, according to their pair wise AS values (see figure 1). We checked manually at least 5% of files from every group, in order to estimate the share of correctly aligned paragraphs in both corpora. Figure 1 shows the estimated percentages of correct alignments in both corpora, dependent on AS. We can see that in case of dissimilar alignments, the UT corpus version is better. The figure shows average values; it does not mean that the alignment quality of a text from the UT corpus was always better than that of the corresponding text from the JRC Vanilla corpus.

Based on the manual evaluation we can say that 95% of all the parallel units in the UT corpus are correct; for JRC Vanilla corpus, the figure is 84%.

We can see from Figure 1 that the lower AS between JRC Vanilla and UT corpus, the worse the alignment quality of the JRC Vanilla version is. In other words, in case of disagreement between alignments, in most cases it would be wise to trust the UT version.

The next question would be: is there a way to estimate the correctness of parallel units without comparing them with another corpus?

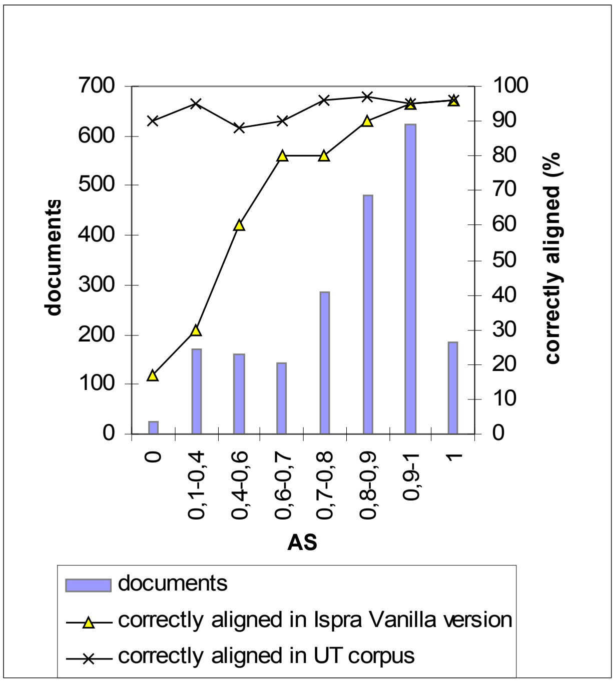

Figure 1. The percentage of correct alignments and number of documents for different values of AS

0-alignments as Indicators of Alignment Quality

We computed the covariation and correlation between the estimated correctness of alignments and the share of 0-alignments in a file.

For UT corpus, the covariation was 0.33 and the correlation 0.02, i.e. the number of 0-alignments and estimated alignment correctness are not interrelated. The reason for this may be the alignment process itself which deleted the 0-alignments of the previous stages, before starting to align smaller units.

For JRC Vanilla corpus, the covariation was -53.82 and the correlation -0.42, i.e. the 0-alignments and estimated correctness are interrelated (although not very strongly). Thus the share of 0-alignments can be used as an indicator of alignment correctness for JRC Vanilla.

It appears that if a length-based aligning algorithm fails to find a matching translation for one paragraph, then this indicates that there is some mismatch between the files altogether and the matching paragraphs found are likely incorrect as well.

If one wants to select some aligned texts from the JRC Vanilla corpus, and has to balance the resulting corpus size and quality, he might base his judgment on the proportion of 0-alignments in files. Figure 2 shows how the corpus characteristics depend on the maximally allowed percentage of 0-alignments.

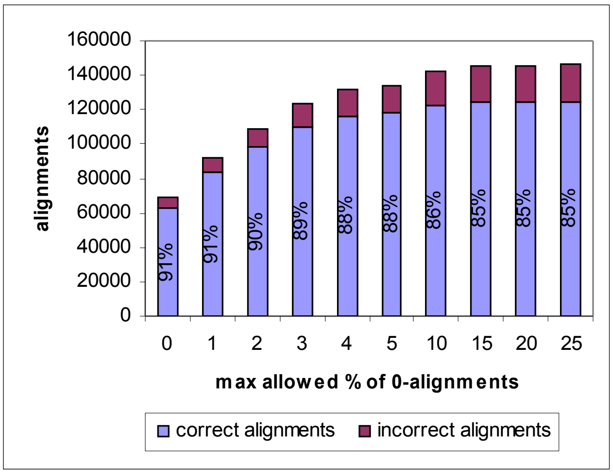

Figure 2. The size and alignment correctness of the Ispra Vanilla version, depending on the maximally allowed % of 0-alignments in a file

Comparison of the JRC HunAlign and UT Corpus

The AS of JRC HunAlign version and UT corpus was larger than AS of JRC Vanilla and UT corpus.

We found that if HunAlign had established a 0-alignment, then this parallel unit and one preceding or following it were incorrect. In 95% of the texts that we checked manually, the alignment outside these 0-alignment regions was completely correct. However, the proportion of 0-alignements is so large in the JRC HunAlign version that the total estimated correctness of the parallel units is 94%, a figure that nevertheless practically equals the 95% of the UT corpus.

Comparison of the JRC Vanilla and HunAlign Versions

One fourth of the texts in these corpora had exactly the same alignments. For the rest of the texts, it appeared that the HunAlign version was always more correct than the Vanilla one.

Results

Evaluations of aligning methods have reached the conclusion that at present it is impossible to single out one method that would be the best for every corpus. It has been found also that assuming that the precision and recall of an alignment method, calculated on the basis of one corpus, is applicable on another corpus, can be misleading (Rosen 2005), (Singh & Husain 2005). Our comparisons confirmed these views. In addition, we found that even if the same method (Vanilla) is used, the results may differ dramatically, if the corpora have been normalized differently or the method applied in a slightly different way.

Our evaluation showed that the HunAlign version of the JRC-Acquis corpus is considerably better than the Vanilla version. We can assert also that in the HunAlign version, a 0-alignment and one preceding or following it are very likely incorrect, and that deleting them would raise the average correctness of the remaining parallel units to 99%.

There are some simple considerations, following of which would enable one to diminish the amount of alignment errors, irrespective of the method used. These considerations were followed while creating the alignments of the UT corpus.

1. One should not trust the source and translation texts to have exactly the same contents and structures. One of the texts may have some parts missing, so the alignment should start from the largest possible blocks for aligning (e.g. parts, chapters) and move step by step towards smaller units.
2. Automatically aligned, but clearly non-parallel blocks should be deleted, because aligning their inner units would yield meaningless results.

The method used for determining AS (UNIX *diff* followed by an additional comparison using Levenshtein distance) can be used also for selecting the most trustworthy alignments.

The comparison methods we used might be suitable for other corpora also, given they have similar partially overlapping counterparts. E.g. the English-Slovenian corpora of legislative texts SVEZ-IJS and JRC-Acquis (Erjavec 2006) may be worth considering.

Conclusions

While it has been common practice to compare and evaluate alignment quality when describing alignment methods, to our knowledge this is the first time when the alignments of completed parallel corpora are compared and evaluated.

It appears that corpora that have been created independently, containing essentially the same texts from independent sources, and which have been aligned with different methods, can be used for evaluating the alignment quality of the corpora themselves.

It appeared that the corpora compared were different not only in their translation versions and alignments, but in their source text parts as well. There was not a single text, the English part of which had completely coincided with that of the corresponding one from the other corpus.

When comparing the corpora, we used the alignment similarity measure AS and this allowed us to economize on manual evaluation.

The percentage of correctly aligned paragraphs ranged from 84% in JRC-Acquis Vanilla version to 94-95% in JRC-Acquis HunAlign version and the UT corpus.

Making use of anchor points in addition to length-based alignment methods proved to increase the correctness by 10% (this is the difference between JRC Vanilla version and JRC HunAlign or UT corpus) even for the short, formally similar texts that are characteristic of EU legislation. The anchor points were used either as an integral part of the alignment process (HunAlign), or for filtering the intermediate alignment results (UT corpus).

The different levels of alignment quality may explain some of the differences in results by (Fishel et al, 2007), observed in statistical MT experiments conducted on different corpora.

The method used for determining AS can be used also for selecting the most trustworthy alignments from the combination of two corpora.

The features that predict the quality of alignments – the proportion of 0-alignments in case of JRC Vanilla, and the mere existence of 0-alignments in case of JRC HunAlign – allow one to select the aligned units that are more trustworthy, even in the absence of a comparable text from another corpus.

The evaluation results of JRC-Acquis corpus might be transferable to other language pairs of the corpus; this assertion needs further investigation, however.

References

Bojar, O. & Žabokrtský, Z. (2006). CzEng: Czech-English Parallel Corpus, Release version 0.5. Prague Bulletin of Mathematical Linguistics, 86. (in print).

Erjavec, T (2006). The English-Slovene ACQUIS corpus. In Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006, Genoa, Italy

Fishel, M., Kaalep, H-J., Muischnek, K. (2007). Estonian-English Statistical Machine Translation: the First Results. In NODALIDA 2007 Conference Proceedings, Tartu, http://dspace.utlib.ee/dspace/bitstream/10062/2589/1/post-Fishel-13.pdf

Gale, W. A. & Church, K. W. (1993) Program for aligning sentences in bilingual corpora. Computational Linguistics 19, 75-102

Koehn, P. (2002) Europarl: A Multilingual Corpus for Evaluation of Machine Translation, Draft, Unpublished, http://people.csail.mit.edu/~koehn/publications/europarl.ps

Langlais, P., Simard, M., & Véronis, J. (1998). Methods and Practical Issues in Evaluating Alignment Techniques. In Joint 17th International Conference on Computational Linguistics (COLING'98) and 36th Annual Meeting of the Association for Computational Linguistics (ACL'98) (pp. 10-14). Montréal.

Rosen, A (2005). In Search of the Best Method for Sentence Alignment in Parallel Texts. In Proceedings of SLOVKO 2005, the Third International Seminar on Computer Treatment of Slavic and East European Languages, VEDA, Bratislava

Samy, D., Sandoval, A.M., Guirao, J.M., Alfonseca, E. (2006) Building a Parallel Multilingual Corpus (Arabic-Spanish-English). In Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006, Genoa, Italy

Singh, A. K. & Husain, S. (2005). Comparison, selection and use of sentence alignment algorithms for new language pairs. In Proceedings of the ACL Workshop on Building and Using Parallel Texts, (pp 99–106), Ann Arbor, Michigan. Association for Computational Linguistics.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006, Genoa, Italy.

Tiedemann, J. & Nygaard, L. (2004). The OPUS corpus – parallel & free. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal.

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón (2005): Parallel corpora for medium density languages. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria.