Czech-Slovak Parallel Corpora for MT between Closely Related Languages*

Petra Galuščáková and Ondřej Bojar

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Abstract. The paper describes suitable sources for creating Czech-Slovak parallel corpora, including our procedure of creating plain text parallel corpora from various data sources. We attempt to address the pros and cons of various types of data sources, especially when they are used in machine translation. Some results of machine translation from Czech to Slovak based on the acquired corpora are also given.

1 Introduction

The Czech language has twice as many users as the Slovak language, resulting in more foreign texts being translated into Czech than Slovak. Czech and Slovak are closely related languages and thus machine translation from Czech to Slovak is a much easier task than translation from a third language to Slovak. If we need to translate some texts from e.g. English to Slovak and these texts are already translated into Czech, it is easier to translate these translations into Slovak.

Depending on the type of machine translation system chosen, large Czech-Slovak parallel corpora may be needed. In any case, such a parallel corpus serves as a good evaluation set.

In the following, various possible sources for acquiring Czech-Slovak parallel corpora are covered. We attempt to describe the pros and cons of each source, especially with respect to the task of training or evaluating MT systems. The subsequent sections are aimed at the usage of the corpora. We describe experiments that we performed with "Moses", a statistical machine translation system that was trained and tuned with the acquired corpora.

2 Tools

Our ultimate goal was to acquire plain text aligned Czech-Slovak sentences. Therefore, the data that we collected required processing. The first step was segmentation into sentences. We used a trainable tokenizer by Ondřej Bojar (Klyueva N., Bojar O. [2]) and adapted it to Slovak for our purposes.

* The work on this project was supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), GAČR P406/10/P259, and MSM 0021620838.

The alignment between Czech and Slovak sentences was found using Hunalign software (Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. [5]).

High quality segmentation is very important for good alignment. Specifically, it is very important that segmentation works in the same manner for both the Czech and Slovak. For example, if there is a sentence break after an ordinal number and a dot in one language, there should be a corresponding sentence break in the second language as well. Mismatches in sentence segmentation lead to output as illustrated in Figure 1. Instead of 1-1 alignment, Hunalign resorts to 2-1 or 1-2 alignments. While this does not completely disqualify such alignments, these non-matching alignments are (in other cases) often of lower quality and are simply removed in subsequent steps. Thus the mismatch in segmentation together with simplistic subsequent filtering leads to unnecessary data loss.

Alignment Type	Czech Sentence	Slovak Sentence
2-1	— <s> Viktor nevnímal hovor a zmatek ve vagónu.	Viktor nevnímal vravu a zmätok vo vagóne.
2-1	"Pryč ode mne, vy zloto! <s> Co vám udělaly ty kačátka?	„Preč odo mňa, vy lotri! čo vám urobili tie kačičky?
1-2	Stáří 23 let. Zoolingvistka.	Vek dvadsaťtri rokov. <s> Zoolingvistka.
1-2	II/ MODLITBA	II <s> MODLITBA

Fig. 1. Several examples of mismatched sentence segmentation leading to non-1-1 alignments. Sentence breaks that required rejoining to achieve the alignment are displayed as "<s>".

3 Sources for Czech-Slovak Parallel Corpora

We surveyed several sources, of parallel Czech-Slovak data. The sources differed in several ways. Some sources were more useful than others due to the ease of extracting aligned data from them. We sought plain text parallel corpora, that required no manual annotation.

- **Books**
  Books in general are a very good source of data, especially for machine translation purposes, thanks to their high quality text and translation. On the other hand, acquiring such data is quite complicated. The Slovak Academy of Science[1] is currently preparing the Czech-Slovak parallel corpus, which is based on books. The use of this corpus is limited due to copyright restrictions. We used an older version of this corpus that contained 118 books in total: 61 Slovak books translated to Czech, 55 Czech books translated

[1] http://korpus.juls.savba.sk

to Slovak and two books translated from a third language into both Czech and Slovak. This version of corpora did not contain alignment, therefore we performed the alignment ourselves.

Books are quite difficult to align, because they often consist of long contiguous texts without reference points. Therefore the quality of automatic alignment needs to be controlled. We also found other problems with aligning. Sometimes the translated text was truncated, more often several sentences were compressed into a single sentence, and in several cases whole passages of text were omitted.

– **Acquis JRC**

Acquis[2] is a parallel corpus created from texts of European Union, which is freely available. This corpus offers large amounts of parallel data for all pairs of official EU languages including Czech and Slovak. Czech and Slovak texts were created from the translation of a third language, English, in most cases. The main drawback of the Acquis texts is their monotonous nature with large portions of texts often being repeated. This problem is illustrated in Table 1 where the number of all sentences (lines after our sentence segmentation) and the number of unique lines are compared.

Source	Lines Total	Lines Unique	%
Acquis CZ	926082	608086	65.66
Acquis SK	926082	632916	68.34
Books CZ	153478	148705	96.89
Books SK	153478	149152	97.18

Table 1. Comparison of the number of all lines (i.e. sentences in our segmentation) to the number of unique lines. Ec-Europa corpus was already deduplicated and therefore is not listed in the table.

Due to the many duplicated sentences, a random subset of the Acquis corpus selected as a test corpus may often contain sentences that are verbatim present also in the remaining "training" data. The results of the (automatic) MT evaluation based on this corpus can thus be overly optimistic if the MT system is trained on this corpus. Another problem is that this corpus is a collection of legislative texts and the vocabulary is somewhat restricted. For these reasons, evaluation based on this corpus cannot be compared to the evaluation based e.g. on books or newspaper articles. Thus, Acquis is a very good source for training data but it should be combined with more disparate sources for the purposes of testing.

[2] http://optima.jrc.it/Acquis

– **Ec-Europa-Eu**
Another source that we examined was the website of the European Commision[3]. This website consists of pages in various language mutations, including Czech and Slovak. Sites in various languages differ by the suffix used in their respective URL and pages with the same name should contain the same text. Thus, alignment at the document level is straightforward.
These texts were manually translated, probably from English into other languages. Unfortunately, very often a portion of a page has been left untranslated but nevertheless is presented under the target language label, so Czech and Slovak pages often contain English parts.
We implemented a custom web crawler for downloading these pages. For technical reasons, we downloaded only a subset of all available pages of the site. In total, we downloaded 25737 Czech and 25918 Slovak web pages. The downloaded pages required some initial cleanup work. We removed all HTML tags and corrected the character encoding. Web pages in Czech and Slovak were paired with each other based on their URLs and the parallel ones were segmented into sentences. Duplicated sentences were removed from these web pages afterwards. We decided to run sentence deduplication before alignment because of the amount of (identical) English text inside both Czech and Slovak variants of the page. The remaining sentences were automatically aligned.

– **Eur-LEX**
The Official Journal of the European Union[4] may be used as another source. This source offers a huge quantity of data. The data is somewhat similar to the Acquis corpora and so similar problems may be associated with it. Here Czech and Slovak documents were also created as a translations from English. Documents in the corpus are in XML format that first required conversion into plain text.
Based on our observation of the document collection, we sorted documents into two types: lists and texts. Sorting was performed automatically based on the average number of words per line. Documents in which the average number of words per line was less than 2.8 were marked as lists. The remaining documents were marked as texts. Next we counted the number of lines in the list documents. If a given document was marked as a list in both Czech and Slovak and it contained the same number of lines in both the Czech and Slovak versions, then this document was marked as a parallel list. This sorting to texts, lists and parallel lists is not very precise, but it proved to be sufficient in most cases. Parallel lists were then aligned line-by-line. Non-parallel lists and texts were aligned by Hunalign software.
The official alignment performed by the publisher of this corpora is expected to be completed in the near future.

[3] http://ec.europa.eu/
[4] http://eur-lex.europa.eu/JOIndex.do

– **Other Possible Sources**
Among other sources, we also translated several sentences from WMT[5]. In the future more web pages from the European Union could be used as data sources. Articles from Project Syndicate[6] are sometimes also translated into Slovak, although they are not generally available on the project web page. If there also exists a Czech version of these articles, they could be used as another source. Sometimes, it is also possible to find news from the Czech News Agency translated into Slovak in a Slovak newspaper. This possible source should be further explored. Another possible source could be movie subtitles translated to Czech and to Slovak.

A comparison of the quantities of data acquired from various sources is given in Table 2. Numbers of documents for various source languages are shown in Table 3.

Source	CZ Words	SK Words	CZ Tokens	SK Tokens	Sentences	Documents
Acquis	20.4 mil	20.6 mil	24.3 mil	24.4 mil	926.1 k	20135
Books	6.6 mil	6.6 mil	8.1 mil	8.1 mil	550.6 k	118
Ec-europa	0.4 mil	0.4 mil	0.4 mil	0.4 mil	24.2 k	1493
Total	27.4 mil	27.6 mil	32.8 mil	32.9 mil	1.5 mil	21746

Table 2. Number of acquired words, tokens and sentences from each type of source. The final version of the Eur-LEX corpus has not yet been completed; therefore, we did not include this corpus. We used an older version of corpora created from books than is currently available.

Language	Documents	Sentences
Czech	55	223.6 k
Slovak	61	321.7 k
Other	21630	955.6 k

Table 3. The number of documents for various source languages

4 Usage of the Corpora

The corpora we collected could have wide ranging utilization. We are primarily interested in machine translation from Czech to Slovak. Since we could manage this task, it will be possible to utilize Czech as a pivot language. We could

[5] http://matrix.statmt.org/test_sets/list
[6] http://www.project-syndicate.org/

translate English texts manually to Czech and then use an automatic translation system for translation into other languages that are similar to Czech – for example Polish, Russian or Slovak. Additional thoughts on this concept may could be found in Hric J., Hajič J., Kuboň V. [1].

Parallel data also facilitate the automatic creation of a Czech-Slovak dictionary. Such a dictionary may find further use in automatic translation systems.

5 Czech to Slovak Automatic Translation

We attempted to use some of the acquired corpora for training and testing automatic translation tool Moses[7] [3]. Initially we exclusively used the Acquis corpus; later we also included data collected from books.

Acquis data were sorted into training, tuning and testing data sets according to the same procedures used by Phillipp Koehn in Euro Matrix project[8]. The training set consisted of 926082 sentences and the tuning set consisted of 4107 sentences. We then used books that were manually translated from Czech to Slovak. Alignments were manually checked and only good alignments from the books were used for training and testing. Only 39 books were used in this experiment. A subset of 4000 sentences from the books was randomly chosen as a testing set, another 4000 sentences were randomly chosen as a tuning set and the rest of the corpus (145478 sentences) was used as the training set. We used all of the books when selecting the tuning and testing sets; therefore, it was possible for the vocabulary that was used in the training set to also appear in the testing set. Due to occasional repetition of sentences in the books, some overlap of the test and training sets was also a possibility. Numbers of sentences from the testing set, that also appear in the training set are listed in Table 4.

Moses was first trained with the Acquis training set and tuned with the Acquis tuning set. Next we used Acquis as a training set and books for tuning. As a third procedure we used books as a traning set and Acquis for tuning and as a fourth procedure we used books for both the training and tuning sets. We also tried training on the Acquis corpus merged with books. We used this merged set for training and the Acquis corpus for tuning as a fifth procedure. Finally, we used books for tuning as the last procedure. The test set acquired from books was used for testing in all of the cases. The result of the evaluation can be found in Table 5. A "BLEU" score [4] was used for the automatic evaluation. This metric is based on a comparison of translated segments to a reference translation.

The size of the Acquis training set is much larger than the size of the training set created from books. In spite of this, results acquired when we use books for training are much higher. Size of tuning sets for books are similar to the Acquis corpora. Using books exclusively for the tuning also improves the results. The best result is, not surprisingly, achieved with the training set composed of books and the Acquis corpora when we are tuning with books.

[7] http://www.statmt.org/moses
[8] http://matrix.statmt.org/test_sets/list

Training Corpus	Identical Sentences	%
Acquis SK	7	0.2
Acquis CS	5	0.1
Books CS	142	3.6
Books SK	122	3.1
Acquis+Books CS	143	3.6
Acquis+Books SK	122	3.1

Table 4. Number of sentences in the testing set that also appear in the training set. Corpora were not pre-filtered.

Training/Tuning Corpus	Training Set Sentences	Tuning Set Sentences	BLEU
Acquis/Acquis	708406	3148	0.1808
Acquis/Books	708406	3802	0.2071
Books/Acquis	137027	3148	0.4661
Books/Books	137027	3802	0.4701
Acquis+Books/Acquis	845433	3148	0.4781
Acquis+Books/Books	845433	3802	0.4887

Table 5. BLEU Evaluation of Moses for Czech-to-Slovak using various data sources. The test set was the same for all cases and contained 3860 sentences of randomly selected sentences from books. Numbers of sentences were counted after filtering out sentences that contained more than 40 tokens.

The difference between highest and lowest scores is very large. This may be caused by there being a wider range of word forms than are used in books but not as wide a range as in the monotonous EU legislation.

However, we are aware of the fact that we tested with the same books as were used for training (despite the disjoint subset of sentences). The vocabulary in the training and test sets, including e.g. proper names, can be thus unnaturally similar. To obtain a more realistic estimate of MT quality, we plan to test using sentences from new books that are not included in the training data.

6 Summary

We described various types of data sources for parallel Czech-Slovak corpora. The initial cleanup of these sources and the necessary steps used to create our parallel corpora were also described.

Additionally, we have given some preliminary results of our machine translation based on the acquired corpora. The results are closely related to the described characteristics of the data sources that we used. We observed a sharp increase in (automatically estimated) MT quality when books were included in the training data. The exact explanation for this has yet to be determined.

Bibliography

[1] Hric J., Hajič J., Kuboň V. (2000). Machine Translation of Very Close Languages. *Proceedings of the 6th Applied Natural Language Processing Conference*, pages 7–12.

[2] Klyueva N., Bojar O. (2008). UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proceedings of International Conference Corpus Linguistics*, pages 188–195.

[3] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

[4] Papineni K., Roukos S., Ward T, Zhy W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

[5] Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. (2005). Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 590–596.