Evaluating Utility of Data Sources in a Large Parallel Czech-English Corpus CzEng 0.9

Ondřej Bojar, Adam Liška, Zdeněk Žabokrtský

Charles University in Prague
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha 1, Czech Republic
{bojar,zabokrtsky}@ufal.mff.cuni.cz, adam.liska@gmail.com

Abstract

CzEng 0.9 is the third release of a large parallel corpus of Czech and English. For the current release, CzEng was extended by significant amount of texts from various types of sources, including parallel web pages, electronically available books and subtitles. This paper describes and evaluates filtering techniques employed in the process in order to avoid misaligned or otherwise damaged parallel sentences in the collection. We estimate the precision and recall of two sets of filters. The first set was used to process the data before their inclusion into CzEng. The filters from the second set were newly created to improve the filtering process for future releases of CzEng. Given the overall amount and variance of sources of the data, our experiments illustrate the utility of parallel data sources with respect to extractable parallel segments. As a similar behaviour can be expected for other language pairs, our results can be interpreted as guidelines indicating which sources should other researchers exploit first.

1. Introduction

Parallel corpora are essential for the training of (statistical) machine translation (MT) systems and used in other NLP tasks as well, e.g. language learning tools or terminology extraction. The paper accompanying the previous release of CzEng 0.7 (Bojar et al., 2008a), confirmed that larger datasets usually improve the quality of MT, even if the additional data are out of the translated domain.

CzEng 0.9 extends the previous release by adding data from several large sources like e-books, parallel web pages and technical documentation. Moreover, CzEng 0.9 has been automatically processed by TectoMT (Žabokrtský et al., 2008) so the whole corpus now includes Czech and English automatic analyses at the morphological, analytical (surface syntactic, labelled "a-" in the sequel) and tectogrammatical (deep syntactic, labelled "t-") layers of description, following Functional Generative Description (Sgall, 1967; Sgall et al., 1986) and the Prague Dependency Treebank (Hajič et al., 2006).

Prior to the release, CzEng 0.9 was automatically checked by various heuristical filters in order to avoid mis-aligned or otherwise malformed parallel sentences, see Bojar and Žabokrtský (2009). This paper describes the filtering process in a closer detail, evaluates the accuracy of the various filters and perhaps most importantly, provides an estimation of utility of individual data sources used in CzEng 0.9.

2. CzEng Data and Processing

This section gives an overview of all types of parallel text resources exploited in CzEng 0.9. The corpus is not claimed to be intentionally balanced in any sense—it simply contains as much material as possible. However, the set of covered topics is quite broad, with style ranging from formal language of laws and technical documents through prose fiction and journalistic language to colloquial language as often appearing in movies. CzEng 0.9 contains exclusively texts that were already publicly available in an electronic form, in most cases downloadable from the Internet.

The proportions of the individual types of texts, which are included into CzEng 0.9, are roughly illustrated in Figure 1.

2.1. Brief Description of Data Sources

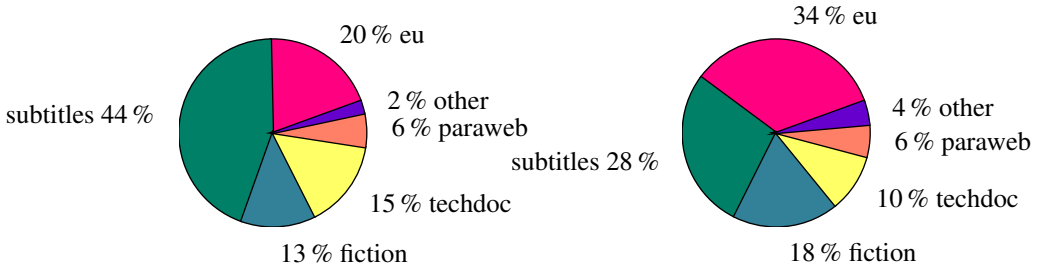
Movie and Series Subtitles (`subtitles`) are being created by a large community of movie fans. While it is not always straightforward to identify parallel versions of subtitles (see Bojar and Žabokrtský (2009)), the abundance of the data makes subtitles an important source even if some pairs remain unexploited.

Unfortunately, the quality of texts available in subtitles is rather unstable. Some authors systematically write "I'II" instead of "I'll", some others leave long passages untranslated, disregard punctuation, or disregard Czech diacritics, etc. The processing pipeline used in CzEng 0.9 tried to remove many of such dubious documents, however from a preliminary inspection of the final corpus, it seems that many errors were not spotted.

Parallel Web Pages (`paraweb`) were identified on the basis of a typical language tag appearing in the URL of a page.[1] If there were two versions of the URL differing in the language tag only, the documents were deemed parallel. Admittedly, this severely restricts the potential of the method as many web sites translate also the URLs in order to rank higher in search engines. No filtering of the pages to avoid e.g. advertisements was performed, only duplicated segments were partially removed.[2] The de-duplication dramatically

[1] The original list of URLs of pages hosted in the `.cz` domain was provided by a Czech search engine.

[2] To reduce the distortion of data distribution, a context-sensitive de-duplication is performed instead of the simple "`sort | uniq`" command: a sliding window of 3 consecutive lines is used and the lines in the window are printed only if no such win-



(a) Sentences (b) Tokens

Figure 1: Types of parallel texts in CzEng 0.9. The depicted proportions are derived (a) from the number of included 1-1 sentence pairs, and (b) from the number of tokens (words and punctuation marks, summed for both languages).

reduces the actual number of segments: from 17M parallel segments (allowing one-to-many and many-to-one mappings) in the originally downloaded data, we are left with only 1.5M parallel segments after the de-duplication.

Fiction (`fiction`) mostly contains electronically available books obtained from various sources, e.g. Project Gutenberg[3] for English and Palmknihy[4] for Czech. The books were aligned based on the author's name and the title and the output of the heuristical aligner was manually checked. Moreover, only documents with reasonable sentence-level alignments were retained.

European Union Law (`eu`) texts come mostly from the freely available corpus JRC-Acquis (Ralf et al., 2006).[5]

Technical documentation (`techdoc`) combines manuals and localization strings of open source projects GNOME and KDE with Microsoft glossaries. Frequently, the segments contain special formatting or substitution codes, although not unified enough to be removed fully automatically.

News texts (`news`) contain articles published regularly on Project Syndicate[6] website, the texts translated from Wall Street Journal for the purposes of Prague Czech-English Dependency Treebank (Cuřín et al., 2004) and also a small but growing collection of news published in Czech and English on the server Aktualne.[7]

User-Contributed Translation at Project Navajo (`navajo`)[8] which provides texts from the English Wikipedia machine-translated to Czech. Users of Navajo are allowed to correct the MT output and these corrections are then available along with the original text. As evaluated in Bojar et al. (2008b), about 70% of contributed segment pairs are of reasonable quality.

2.2. Common Processing Pipeline

All the documents in CzEng 0.9 were converted from many source formats to UTF-8 encoded plain text, segmented into sentences using a trainable tokenizer[9] and sentence-aligned using Hunalign (Varga et al., 2005).

Only 1-1 aligned sentences (about 82% of segment pairs) were further considered, although some 2-1 and 1-2 segments seem to indicate a missing sentence break in one of the languages and could be re-segmented to obtain two 1-1 aligned pairs.

Sentence-aligned plaintext format was used to perform checks to filter out either mis-aligned or simply bad segments. If any of the filters marks the segment as inappropriate, the segment is removed and is not included in the corpus.

Additional filters were subsequently developed which used the CzEng 0.9 Export Format on input and did not affect the final corpus.

3. Filters Included in CzEng 0.9

Filters implemented prior to the release of CzEng 0.9 included the following checks:

- the Czech and English sentences are identical strings (usually untranslated text from a website),
- the lengths of the sentences are too different (usually due to a wrong alignment or a wrong sentence segmentation),
- there is no Czech word on the Czech side or English word on the English side according to case-insensitive wordlists from the Czech and British National Corpora, [10]

dow was printed before. For instance, for the lines "a b c a b c b d b" we get "a b c b d b". The second occurrence of "a b c" got removed but the overall distribution of "b" is influenced less.

[3] `http://www.gutenberg.org/`
[4] `http://www.palmknihy.cz`
[5] `http://wt.jrc.it/lt/acquis/`
[6] `http://www.project-syndicate.org/`
[7] `http://aktualne.centrum.cz/czechnews/`
[8] `http://www.navajo.cz/`
[9] `http://ufal.mff.cuni.cz/euromatrixplus/`
[10] Longer words are preferred for the test: if there are some words longer than three letters, at least one of them has to be confirmed in the word list. If all words contain at most three letters, also shorter words are accepted for the word list check.

- there is a suspicious character (either a non-printable one or an unlikely symbol) or a repeating sequence of a character.
- clearly suspicious segmentation or tokenization: letters interleaved by spaces, academic titles appearing at the end of the sentence instead of being followed by a name,
- outstanding HTML entities or tags (all entities and tags should have been interpreted or removed during the conversion to plain text),
- relicts of metainformation, e.g. Project Gutenberg headers, EU legislation headers, lines containing only file pathnames.

4. Additional Filters

Additional filters were implemented in order to complement the previous list and improve the performance of the filtering process for future releases. Among these additional filters were for example the following:

- the English sentence contains non-ASCII characters that are not present in the Czech sentence,
- the use of numbers in the Czech and English sentences is different (number filter),
- word-alignment score of each sentence pair is below a given threshold.

In future, we also plan to implement filters based on the probability of the sentences as scored e.g. by:

- an n-gram language model, possibly trained on morphological tags instead of word forms to increase robustness,
- a dependency-based language that would evaluate whether the dependency edges of the sentence are similar to the edges appearing in a good quality text or a manually annotated treebank.

4.1. Non-ASCII Characters in English

A great number of misalignments were caused by pseudo-parallel web pages where the English translation contained untranslated Czech text. To filter these out, the ASCII filter marked as incorrect those alignments where the English sentence contained non-ASCII characters that were not present in the Czech sentence. Some special characters such as quotation marks or dashes were disregarded.

4.2. Number Filter

Another method developed compares the use of numbers in a sentence pair. We implemented a filter that for all numbers in the English sentence tries to find an equivalent in the Czech sentence, i.e. either the same number or a possible translation.

Often, a numeral in an English sentence would be expressed in words in the Czech sentence. The filter generates a list of word equivalents for which it looks in the Czech sentence. For example, for the number 21, there are at least four different alternatives besides the numerical expression: "dvacet jedna", "dvacetjedna", "jednadvacet", "jedenadvacet".

4.3. Word-Alignment Score

The last filtering method employed is based on word-alignment probabilities produced by GIZA++ toolkit (Och and Ney, 2003). A similar approach was described in (Khadivi and Ney, 2005).

For this purpose, Hidden Markov Model, IBM Model 1, IBM Model 3 and IBM Model 4 (in this order) were trained on lemmas and alignment scores were obtained for both directions. The final alignment score for each pair, which was then compared to a given threshold, was calculated by the following equation:

$$\begin{aligned} Score\left(e_1^J, f_1^I\right) = \frac{1}{J}\log\left(p\left(\boldsymbol{e}, a \mid \boldsymbol{f}\right)\right) \\ + \frac{1}{I}\log\left(p\left(\boldsymbol{f}, a \mid \boldsymbol{e}\right)\right) \end{aligned} \quad (1)$$

5. Manual Evaluation of Filters

To evaluate individual filters, we randomly selected two sets of 1000 sentence pairs. The first set was used to evaluate filters included in CzEng 0.9 and was thus selected from the aligned plaintext files just before the application of the filters described in Section 3.. Note that we included only 1-1 aligned segments in the evaluation.

The second set was taken from the files publicly released as CzEng 0.9. Because the public release of CzEng 0.9 has been already randomized (at the level of short sequences of sentences), we simply used the first 1000 segments from the section `30train`. This set was used to evaluate newly implemented filters. (The filters operate on the "export format" of CzEng 0.9, which is somewhat richer plaintext.)

For both of these sets, we evaluate the overall precision and recall:

$$\text{overall recall} = \frac{\left|\begin{array}{c}\text{segs. marked by both human} \\ \text{and at least one filter}\end{array}\right|}{|\text{segs. marked by human}|} \quad (2)$$

$$\text{overall precision} = \frac{\left|\begin{array}{c}\text{segs. marked by both human} \\ \text{and at least one filter}\end{array}\right|}{|\text{segs. marked by at least one filter}|} \quad (3)$$

The recall of individual filters is evaluated on the 1000 random segments using the formula:

$$\text{recall of filter } f = \frac{\left|\begin{array}{c}\text{segs. marked by both human} \\ \text{and filter } f\end{array}\right|}{|\text{segs. marked by human}|} \quad (4)$$

Note that in this evaluation, we do not aim at the recall of 100% for individual filters but rather at the overall recall of 100%. Some filters are very specific and their recall is expected to be low.

	Precision	Recall
Not enough letters	94%	7%
Mismatching lengths	91%	11%
Repeated character	88%	2%
No English word	80%	11%
Suspicious char.	75%	1%
Identical	72%	26%
No Czech word	67%	2%
Too long sentence	12%	0%
Extra header	2%	0%
Overall (all filters)	57%	42%
Overall (evaluated filters only)	57%	41%

Table 1: Manual evaluation of CzEng 0.9 filters. Only filters that apply most often were evaluated for precision.

Within the 1000 segments data set, some filters fired only e.g. three times. In order to evaluate the precision of individual filters reliably, we had to extend the set of manually annotated segments. For each of the evaluated filters, we selected 200 (for CzEng filters) or 500 (for additional filters) segments where the filter fired. The precision reported for individual filters has been estimated on the correspoding 200 or 500 segment dataset using the following formula. The S denotes the dataset size (200 or 500).

$$\text{precision of filter } f = \frac{\left|\begin{array}{c}\text{segs. marked}\\ \text{by the human annotator}\end{array}\right|}{S} \qquad (5)$$

5.1. Evaluation of CzEng Filters

Table 1 documents that the original CzEng filters are not very reliable. Their overall precision falls below 60%. The "Overall" figures provide the precision and recall for the whole ensemble of filters (either all implemented, or all evaluated): if any of the filters fired, the segment is deemed invalid. We see that the ensemble of evaluated filters indeed does most of the work, the differences in both precision and recall between the evaluated and the full ensemble are negligible.

One of the least precise filters is the one that marks all sentences over 400 words as wrong, although the majority of these segments seems fine according to the manual evaluation. While it seems a pity to lose all word translations available in such long sentences, we realize that e.g. the word alignment tools are usually not capable of handling this long sentences anyway so e.g. machine translation using CzEng 0.9 data is not influenced much by this limitation.

The worst evaluated filter is the removal of EU legislation metainformation (Extra header). In most cases, the filter removes valid but rather uninformative segments "Article 123"="Článek 123" but a few cases, the actual title of the article got removed as well.

The situation is rather difficult for the filter searching for non-translated sentences (Identical). This filter reached the highest recall but its precision is disputable. In some cases, it is absolutely correct to preserve exactly the original sentence, e.g. when it contains only an author's name. Again, we believe a machine translation system would not lose much if not trained on identical segments as most MT systems preserve input words if unknown.

Filter	Precision	Recall
Non-ASCII characters in English	100%	4%
Number	88%	6%
Word-alignment scores	77%	33%
Overall	79%	40%

Table 2: Manual evaluation of additional error-detection filters.

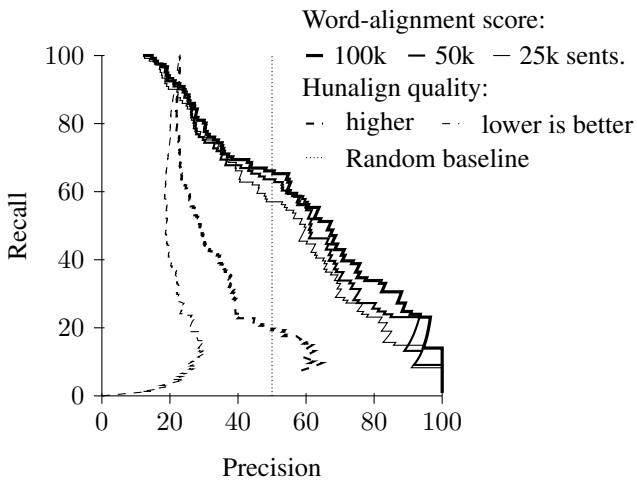


Figure 2: ROC curves for filtering based on sentence alignment quality (Hunalign) and word alignment scores.

5.2. Evaluation of Additional Filters

Table 2 illustrates that the newly implemented filters have about the same recall (note that we run the new filters only on segments that already passed the original CzEng filtering) and precision approaching 80%. The most reliable is the filter against non-ASCII characters in English while word-alignment scores reveal their stochastic nature with decreased reliability.

5.3. ROC Curves for Alignment Filters

The precision and recall of the filter based on word-alignment score can be tuned by choosing the threshold value. In Figure 2, we plot the precision against the recall for many different thresholds. The random baseline achieves the precision of 50% for all recalls.

By setting the threshold for word-alignment score (Eq. 1) at -10, we achive the precision of 77% and the recall of 33% when the word alignment was trained on 100k sentences. Reducing the size of GIZA++ training data reduces the reliability of this filter.

Currently, we are satisfied with the mentioned precision-recall balance but with more filters in the ensemble we would prefer higher precision, assuming that other filters will help to reach some satisfactory recall.

The other two curves in Figure 2 are based on sentence alignment quality for individual segments, as reported by

Bad 1-1 Segments [%]		Most Frequent Errors
subtitles	4.6	Mismatching lengths (42.0%), Identical (27.3%), No English word (10.9%),
eu	33.3	Identical (39.9%), No English word (19.2%), Not enough letters (17.2%),
techdoc	10.2	Identical (37.9%), No English word (28.4%), Not enough letters (10.0%),
paraweb	59.5	Identical (61.7%), No English word (25.1%), Mismatching lengths (3.3%),
fiction	3.1	Mismatching lengths (54.9%), Suspicious char. (14.6%), Repeated character (6.1%),
news	3.8	Identical (54.1%), Suspicious char. (17.7%), No English word (9.3%),
navajo	11.9	Identical (40.9%), No English word (19.0%), Not enough letters (11.7%),

Table 3: Percentage of 1-1 sentence pairs rejected by various error-detection filters.

Bad 1-1 Segments [%]		Most Frequent Errors
subtitles	6.8	Alignment score (94.5%), Number (4.7%), ASCII (2.1%),
eu	3.3	Alignment score (68.7%), Number (37.9%), ASCII (8.4%),
techdoc	3.4	Alignment score (93.7%), Number (9.6%), ASCII (0.4%),
paraweb	17.6	ASCII (51.2%), Alignment score (31.1%), Number (28.1%),
fiction	7.4	Alignment score (86.0%), Number (11.0%), ASCII (5.3%),
news	2.2	Alignment score (55.3%), Number (34.2%), ASCII (23.7%),
navajo	1.9	Alignment score (57.1%), Number (28.6%), ASCII (14.3%),

Table 4: Percentage of 1-1 sentence pairs rejected by additional error-detection filters.

Hunalign (Varga et al., 2005). The sentence alignment quality is in fact a very rough approximation of word-alignment score with a fixed dictionary and considering only the top-scoring translation of a word. We see that in our case, the sentence alignment quality is a very poor indicator of the quality of individual segments. Given the low performance, we suspected we are mis-interpreting the numbers reported by Hunalign assuming higher means better. Therefore we also plot the ROC curve if lower meant better. During the process of sentence alignment, the presence of surrounding segments and the aim to reach a reasonable sentence length ratio compensate for the deficiency.

6. Utility of Data Sources

Table 3 displays the percentage of 1-1 aligned sentences from noisy data sources with one or more errors. The second column in the table lists the most frequent error in each of the sections.

While many of the errors could have been corrected in earlier stages of corpus cleaning, the current release of CzEng simply removes such suspicious segments.

The overall most frequent error is "Identical", and we see that e.g. more than 36% of web data (61.7% out of 59.5% of erroneous segments) are removed due to this error. Unfortunately, many of the seemingly parallel web pages contain non-translated sections. The cleanest source is probably the ebooks section with some errors in segmentation or alignment (Mismatching lengths).

Table 4 displays the same statistics for our additional filters. These data were obtained by running the filters on a testset from the public release of CzEng 0.9 corpus. This testset consisted of the first 100,000 sentence pairs from `30train`. The most frequent error in most data sources were low word-alignment scores of aligned sentence pairs, leading to the rejection of 1–6% of segments. Parallel web pages were the only exception, with more than half of the filtered sentences marked by the filter looking for non-ASCII characters in the English sentence.

We also observed that the overlap among the additional filters is rather low.

7. Conclusion

We have presented the filters applied in the cleaning process of texts gathered for the parallel corpus CzEng. The set of filters originally implemented for CzEng 0.9 was extended and all filters were manually evaluated for precision and recall.

Future versions of CzEng will benefit from the presented study: we will attempt at improving the existing low performing filters and we will use all the additionally implemented ones as well.

Given the broad scope of CzEng data sources, we can estimate the utility of individual data types for a parallel corpus. The most reliable sources of parallel sentences are electronic books, subtitles and news with a relatively low number of bad 1-1 segments followed. The worst sources include European legislation and parallel web pages containing very often non-translated segments or other noise.

8. Acknowledgement

The work on this project was supported by the grants MŠMT ČR LC536, GAČR P406/10/P259, EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), and MSM0021620838.

9. References

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng 0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92. in print.

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. 2008a. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of LREC'08*, Marrakech.

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. 2008b. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of LREC'08*, Marrakech.

Jan Cuřín, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. 2004. Prague Czech-English Dependency Treebank, Version 1.0. Linguistics Data Consortium, Catalog No.: LDC2004T25.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0. LDC, Philadelphia.

Shahram Khadivi and Hermann Ney. 2005. Automatic Filtering of Bilingual Corpora for Statistical Machine Translation. In *Natural Language Processing and Information Systems*, volume 3513/2005, pages 263–274.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2142–2147. ELRA.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing RANLP 2005*, pages 590–596, Borovets, Bulgaria.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio. Association for Computational Linguistics.