CzEng: Czech-English Parallel Corpus

Release version 0.5

Ondřej Bojar, Zdeněk Žabokrtský
{bojar,zabokrtsky}@ufal.mff.cuni.cz

Abstract

We introduce CzEng 0.5, a new Czech-English sentence-aligned parallel corpus consisting of around 20 million tokens in either language. The corpus is available on the Internet and can be used under the terms of license agreement for non-commercial educational and research purposes. Besides the description of the corpus, also preliminary results concerning statistical machine translation experiments based on CzEng 0.5 are presented.

1 Introduction

CzEng 0.5[1] is a Czech-English parallel corpus compiled at the Institute of Formal and Applied Linguistics, Charles University, Prague in 2005-2006. The corpus contains no manual annotation. It is limited only to texts which have been already available in an electronic form and which are not protected by authors' rights in the Czech Republic. The main purpose of the corpus is to support Czech-English and English-Czech machine translation research with the necessary data. CzEng 0.5 is available free of charge for educational and research purposes, however, the users should become acquainted with the license agreement.[2]

2 CzEng 0.5 Data

CzEng 0.5 consists of a large set of parallel textual documents mainly from the fields of European law, information technology, and fiction, all of them converted into a uniform XML-based file format and provided with automatic sentence alignment. The corpus contains altogether 7,743 document pairs. Full details on the corpus size are given in Table 1.

2.1 Data Sources

We have used texts from the following publicly available sources:

- Acquis Communautaire Parallel Corpus (Ralf et al., 2006),
- The European Constitution and KDE documentation from corpus OPUS (Tiedemann and Nygaard, 2004),
- Readers' Digest texts were partially made available already in (Čmejrek et al., 2004),
- Kačenka was previously released as (Rambousek et al., 1997); because of the authors' rights, CzEng 0.5 can include only its subset, namely the following books:
  - D. H. Lawrence: Sons and Lovers / Synové a milenci,
  - Ch. Dickens: The Pickwick Papers / Pickwickovci,
  - Ch. Dickens: Oliver Twist,
  - T. Hardy: Jude the Obscure / Neblahý Juda,

[1] http://ufal.mff.cuni.cz/czeng/

[2] http://ufal.mff.cuni.cz/czeng/license.html

  - T. Hardy: Tess of the d'Urbervilles / Tess z d'Urbervillu,
- Other E-books were obtained from various Internet sources; the English side comes mainly from Project Gutenberg.[3] CzEng 0.5 includes these books:
  - Jack London: The Star Rover / Tulák po hvězdách,
  - Franz Kafka: Trial / Proces,
  - E.A. Poe: The Narrative of Arthur Gordon Pym of Nantucket: Dobrodružství A.G.Pyma,
  - E.A. Poe: A Descent into the Maelstrom / Pád do Malströmu,
  - Jerome K. Jerome: Three Men in a Boat / Tři muži ve člunu.

	Document pairs	Sentences		Words+Punctuation	
		Czech	English	Czech	English
Acquis Communautaire	6,272	1,101,610	930,626	14,619,572	16,079,043
	81.0%	77.6%	71.8%	78.9%	76.6%
European Constitution	47	11,506	10,380	138,853	176,096
	0.6%	0.8%	0.8%	0.7%	0.8%
Samples from European Journal	8	5,777	4,993	104,560	133,136
	0.1%	0.4%	0.4%	0.6%	0.6%
Readers' Digest	927	121,203	128,305	1,794,827	2,234,047
	12.0%	8.5%	9.9%	9.7%	10.6%
Kačenka	5	62,696	69,951	1,034,642	1,188,029
	0.1%	4.4%	5.4%	5.6%	5.7%
E-Books	5	17,140	17,495	330,118	399,607
	0.1%	1.2%	1.4%	1.8%	1.9%
KDE	479	98,789	133,897	495,052	784,316
	6.2%	7.0%	10.3%	2.7%	3.7%
Total	7,743	1,418,721	1,295,647	18,517,624	20,994,274
	100.0%	100.0%	100.0%	100.0%	100.0%

Table 1: CzEng 0.5 sections and data sizes.

2.2 Preprocessing

Since the individual sources of parallel texts differ in many aspects, a lot of effort was required to integrate them into a common framework. Depending on the type of the input resource, (some of) the following steps have been applied on the Czech and English documents:

- conversion from PDF, Palm text (PDB DOC), SGML, HTML and other formats,
- encoding conversion (everything converted into UTF-8 character encoding), sometimes manual correction of mis-interpreted character codes,
- removing scanning errors, removing end-of-line hyphens,
- file renaming, directory restructuring,
- sentence segmentation,
- tokenization,
- removing long text segments having no counterpart in the corresponding document,
- adding sentence and token identifiers,
- conversion to a common XML format.

For the sake of simplicity, the tokenization and segmentation rules were reduced to a minimum. This decision leads to some unpleasant differences in tokenization and segmentation compared to the "common standard" of Penn-Treebank-like or Prague-Dependency-Treebank-like annotation.[4]

[3] `http://www.gutenberg.org/`

[4] A different character class (digit, letter, punctuation) always starts a new token. Adjacent punctuation characters are encoded as separate tokens. Consecutive periods (...) thus lead to a sequence of one-token sentences. Moreover, no abbreviations were searched for. This hurts especially with titles (Dr.) or abbreviated names (O. Bojar), because a period followed by an upper-case letter is treated as the sentence boundary. All such expressions are thus split into several sentences.

English-Czech	1-1	0-1	1-2	2-1	1-0	1-3	0-2	3-1	Other
Alignment pairs	924,543	97,929	70,879	69,558	64,490	23,538	8,526	6,768	24,943
	71.6%	7.6%	5.5%	5.4%	5.0%	1.8%	0.7%	0.5%	1.9%

Table 2: Sentence alignment pairs according to number of sentences.

2.3 Sentence Alignment

All the documents were sentence-aligned using the `hunalign` tool[5] (Varga et al., 2005). All the settings were kept default and we did not use any dictionary to bootstrap from. Hunalign collected its own temporary dictionary to improve sentence-level alignments.

The number of alignment pairs according to the number of sentences on the English and Czech side is given in Table 2.

3 First Machine-Translation Results Using CzEng 0.5

To provide a baseline for MT quality, we report BLEU (Papineni et al., 2002) scores of a state-of-the-art phrase-based MT system Moses.[6]

For this experiment, we selected 1-1 aligned sentences up to 50 words from CzEng 0.5. From this subcorpus, a random selection of three independent test sets (3000 sentences each) was kept aside and the remaining 870k sentences were used for training. The training data contained 9.7M Czech and 11.4M English tokens (words and punctuation).

Table 3 reports baseline BLEU scores on 3000-sentence test set with 1 reference translation. The texts were only lowercased (including the reference translation) and no other special preprocessing was performed. No advanced features of Moses such as factored translation were utilized. We ran the experiment three times, always using one of the test sets to tune model parameters, another to evaluate the performance on unseen sentences and ignoring the third test set. For curiosity we also report BLEU scores when not translating at all, i.e. pretending that the source text is a translation in the target language. Only some punctuation, numbers or names thus score.

Our results cannot be compared to previously reported Czech-English machine translation experiments (Čmejrek, Cuřín, and Havelka, 2003; Bojar, Matusov, and Ney, 2006),[7] because those experiments used a different 4 or 5-reference test set consisting of 250 sentences only.

The relatively high scores we have achieved are caused by the nature of our data. Most of our training data come from Acquis Communautaire and contain European legislation texts. Although there should be no reoccuring documents in our collection, there is a significant portion of sentences that repeat verbatim in the texts. Naturally, such frequent sentences can get to the randomly chosen test sets. A check of the three test sets revealed that only 1823±13 sentence pairs did not occur in training data. In other words, more than a third of the sentences in each test set appears already in the training data.

4 Summary And Further Plans

We have presented CzEng 0.5, a collection of Czech-English parallel texts. The corpus of about 20 million tokens is automatically sentence aligned. CzEng 0.5 is available free of charge for educational and research purposes, the licence allows collecting statistical data and making short citations. To our

[5] `http://mokk.bme.hu/resources/hunalign`

[6] Moses has been developed during a summer workshop at Johns Hopkins University, as a drop-in replacement for Pharaoh (Koehn, 2004). See `http://www.clsp.jhu.edu/ws2006/groups/ossmt/` for more details.

[7] English→Czech translation has also been attempted at the JHU workshop, report forthcoming.

	To English	To Czech
Not translating at all	5.98±0.68	5.93±0.67
Baseline Moses translation	42.57±0.55	37.41±0.58

Table 3: BLEU scores of a baseline MT system trained and evaluated on CzEng 0.5 data. Test set of 3000 sentences, 1 reference translation.

knowledge, it is the biggest and the most diverse publicly available parallel corpus for the Czech-English pair.

In the future, we plan to further enlarge CzEng. Even now we are aware of various sources of parallel material available on the Internet and not included in CzEng; however, in most of these cases it seems impossible to make any use of such data without breaking the authors' rights.

Future versions of CzEng will contain (machine) annotation of the data on various levels up to deep syntactic layer. We also plan to designate subsections of CzEng as standard development and evaluation data sets for machine translation, paying proper attention to cleaning up of these sets. Our future plans also include experimenting with several machine translation systems.

5 Acknowledgement

The work on this project was partially supported by the grants GAČR 405/06/0589, GAUK 351/2005, Collegium Informaticum GAČR 201/05/H014, ME838 and MŠMT ČR LC536.

References

Bojar, Ondřej, Evgeny Matusov, and Hermann Ney. 2006. Czech-English Phrase-Based Machine Translation. In *FinTAL 2006*, volume LNAI 4139, pages 214–224, Turku, Finland, August. Springer.

Čmejrek, Martin, Jan Cuřín, and Jiří Havelka. 2003. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90. Association for Computational Linguistics, April.

Čmejrek, Martin, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of LREC 2004*, Lisbon, May 26–28.

Koehn, Philipp. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In Robert E. Frederking and Kathryn Taylor, editors, *AMTA*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124. Springer.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, Pennsylvania.

Ralf, Steinberger, 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.

Rambousek, Jiří, Jana Chamonikolasová, Daniel Mikšík, Dana Šlancarová, and Martin Kalivoda. 1997. KAČENKA (Korpus anglicko-český - elektronický nástroj Katedry anglistiky). http://www.phil.muni.cz/angl/kacenka/kachna.html.

Tiedemann, Jörg and Lars Nygaard. 2004. The OPUS corpus - parallel & free. In *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*, Lisbon, May 26–28.

Varga, Dániel, 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.