A High-Quality Web Corpus of Czech

Johanka Spoustová, Miroslav Spousta

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University Prague, Czech Republic
{johanka,spousta}@ufal.mff.cuni.cz

Abstract

In our paper, we present main results of the Czech grant project Internet as a Language Corpus, whose aim was to build a corpus of Czech web texts and to develop and publicly release related software tools. Our corpus may not be the largest web corpus of Czech, but it maintains very good language quality due to high portion of human work involved in the corpus development process. We describe the corpus contents (2.65 billions of words divided into three parts – 450 millions of words from news and magazines articles, 1 billion of words from blogs, diaries and other non-reviewed literary units, 1.1 billion of words from discussions messages), particular steps of the corpus creation (crawling, HTML and boilerplate removal, near duplicates removal, language filtering) and its automatic language annotation (POS tagging, syntactic parsing). We also describe our software tools being released under an open source license, especially a fast linear-time module for removing near-duplicates on a paragraph level.

Keywords: corpus, web, Czech

1. Introduction

Due to the large expansion of the Internet in the recent years, web space became very rich and valuable mine for language resources of various kind, especially mono- and bilingual text corpora, eg. (Baroni and Kilgarriff, 2006). The aim of our project was to exploit the Czech web space and build a web corpus of Czech, which will be useful both for research in theoretical linguistics and for training NLP applications (machine learning in statistical machine translation, spoken language recognition etc.)

There already exists a large corpus of Czech texts, Czech National Corpus (CNC, 2005), compiled from texts obtained directly from publishers (books, newspapers, magazines etc.), but the legal restrictions do not allow the corpus creators to freely distribute the data.

Generally, due to the author's law, one cannot freely distribute whole texts downloaded from the web neither, but our aim was to find a way how to make our web corpus accessible and downloadable for both professionals and general public, at least in some modified, limited form (see section 7.).

2. Texts selection and the cleaning process

After investigating other possibilities, we have chosen to begin with manually selecting, crawling and cleaning particular web sites with large and good-enough-quality textual content (e.g. news servers, blog sites, young mothers discussion fora etc.). In our selection, we were guided by our knowledge of the Czech Internet and by the results of NetMonitor.cz – a service monitoring Czech web sites popularity and traffic.

For web pages selection and their HTML markup and boilerplate removal, we used manually written scripts for each web site. This approach made us sure, compared to completely automatic cleaning approaches such as (Spousta et al., 2008), that the corpus will contain only the desired content (pure articles, blogs and discussion messages) and we will avoid fundamental duplicates (perexes and samples from articles and blogs, repetitions of first messages on each discussion page in some fora etc.). Additionally, we have removed the documents resulting into empty or nearly empty raw texts from the corpus, including the basic HTML level and the URLs lists.

After downloading and cleaning a few carefully selected sites, we were pleasantly surprised with the size of the acquired data. For example, the poetry server pismak.cz provided us with 40 millions of words[1] of amateur poems, one of the most popular news servers idnes.cz contained 94 millions of words in articles contents, 118 millions of words in articles discussions and 54 millions of words in blogs, and mothers visiting discussion server modrykonik.cz have produced 313 millions of words in their discussions.

For comparison, first version of Czech National Corpus (CNC, 2005), the biggest corpus of Czech, from the year 2000, contained 100 millions of words, and the latest version contains 300 millions of words in balanced texts (fiction, technical literature and news) and 1 billion of words in news texts. All these texts were obtained directly from the publishers and are not available for download from the web.

Encouraged by the size, and also by the quality of the texts acquired from the web, we decided to compile the whole corpus only from particular, carefully selected sites, to proceed the cleaning part in the same, sophisticated manner, and to divide the corpus into three parts – articles (from news, magazines etc.), discussions (mainly standalone discussion fora, but also some comments to the articles in acceptable quality) and blogs (also diaries, stories, poetry, user film reviews). Until now, we have acquired about 3.8 billions of words in raw texts resulting into 2.6 billions of words after near-duplicate detection and language detection, from only about 40 web sites.

[1] sizes of the raw texts, i.e. after HTML markup and boilerplate removal, but before near-duplicate detection and language detection

At the time of writing this article, the total number of Czech top level domains is over 800 000. Naturally, the average number of data obtainable from one site decreases with the decreasing popularity of the site – for example, the most popular Czech blog engine blog.cz provided our corpus with over 1 billion of words (in raw texts), while its competitors, blogspot.com (restricted to Czech texts only), bloguje.cz and sblog.cz contained only 87, 77 and 52 millions of words, respectively.

Table 1 shows the sizes of the parts of the corpus during the downloading and cleaning process. For HTML sources, we show the size of the data in gigabytes. After HTML and boilerplate removal, the data became "raw texts" and the sizes are presented in gigabytes, tokens (words plus punctuation) and words (without punctuation). Next steps (whose resulting sizes are presented in tokens and words) are near-duplicate removal ("deduplicated") and finally, language detection ("cz-only").

3. Near-duplicate detection algorithm

According to our web pages selection and the downloading and cleaning methodology (c.f. section 2.), no duplicates caused by the basic nature of the web (i.e. the same sites under different URLs, the same copyright statements etc.) should appear in our corpus. Still, some near duplicates on the document or paragraph level may appear as parts of the author texts, for example press releases or jokes are being often copied among the different sites or even within the same site.

Thus, we decided to remove the duplicates on paragraph level. One can argue, whether the nature of the documents will not be affected by the gaps caused by removal of some paragraphs. But due to the forms of public distribution of the corpus (N-grams, shuffled sentences, see section 7.) this question becomes irrelevant. Linguists, who will manually investigate the corpus in its original form through our simple query interface, can profit from the links to original websites, incorporated in the query interface.

Back to the technical aspect of the process, there are several different approaches to the duplicate detection task at the document level. In the area of web page near-duplicate detection, the "state-of-the-art" algorithms include (Broder et al., 1997) shingling algorithm and (Charikar, 2002) random projection based approach. The former one may require quadratic number of comparisons of the documents, the later one does not contain an explicit interpretation of similarity.

Our similarity measure is based on n-gram comparison and is easy to interpret: we consider two documents to be similar, if they share at least some number of n-grams.

In order to achieve linear run-time, we take an iterative approach and modify our measure of similarity: we do not compare two documents at a time, instead, we compare document n-grams to all previously added documents. We start with a single document and every time a new document is considered for addition in the corpus, we compute a percentage of n-grams that the document shares with all previously added ones. Using this algorithm, we can continuously expand the corpus size while detecting duplicate documents.

To reduce memory footprint, we store n-grams in a set implemented using the Bloom filter (Bloom, 1970). This data structure stores data very efficiently at the cost of adding a (possibly small) probability of false-positive result. The false-positive rate may be influenced by setting the algorithm parameters, such as number of hashing functions and a target array size.

For the purposes of the Czech Web corpus, we drop paragraphs containing more than 30% seen 8-grams, and we set 1% to be the maximum false-positive rate, which leads to 1.25 bytes used per n-gram. As the number of n-grams corresponds to the number of words, the memory consumed by the deduplication task was about 6 GB and our implementation of the Bloom filter algorithm achieved processing time more than 1 billion tokens per hour (Intel Xeon E5530, 2.4 GHz).

After performing the deduplication algorithm with the described parameters, the corpus size was reduced by about 20 % (see Table 1).

4. Language detection module

Because of historical reasons, a lot of Slovak speakers participate in the Czech web space using their mother tongue (Slovak is very similar to Czech and in general, Czech and Slovak speakers understand each other). In addition, some of us grown up in 7-bit times still use "cestina" instead of "čeština" sometimes, i.e. we omit the diacritics in our written informal communication (email, discussions).

These are main language discrepancies we needed to focus on while developing our language detection module – because of their high frequency in the web data and because of their similarity to original Czech. Indeed, a variety of other languages may also appear in the Czech web space. As our target audience uses both statistical processing and manual inspection, our aim was to leave only fully correct Czech sentences.

Thus, our language filter module consists of two parts: unaccented words („cestina“ and „slovencina“) filter, and a general language filter.[2]

For the first part (filtering unaccented paragraphs), we have developed a detection tool based on frequencies of particular words. We have constructed a list of Czech and Slovak words fulfilling two conditions: 1) they contain at least one accent, and 2) when deaccented, they do not form valid words. Then, we have simply discarded paragraphs (or documents), where the number of such words has exceeded number of accented words. Our aim here was to discard sentences where too many unaccented words were present.

For the second part (language filtering), we have begun with using Google Compact Language Detection Library, part of the Google Chrome browser code, that suggests a translation of web pages. It is based on character 4-grams and supports 52 languages. Although it is compact in size and works well on whole web pages contents, applying it to

[2]It may seem more straightforward to use a general language filter to detect unaccented paragraphs as well, but there is an obstacle in this approach: there are many perfectly correct Czech sentences that do not contain accented words at all and thus a general classifier could not distinguish between unaccented and correct Czech texts.

	articles	discussions	blogs	all
HTML	88 GB	192 GB	109 GB	389 GB
raw text	8.4 GB	16 GB	18 GB	42 GB
raw text (tokens)	737 mil.	2,089 mil.	2,038 mil.	4,864 mil.
raw text (words)	611 mil.	1,674 mil.	1,575 mil.	3,860 mil.
deduplicated (tokens)	634 mil.	1,943 mil.	1,496 mil.	4,073 mil.
deduplicated (words)	531 mil.	1,579 mil.	1,176 mil.	3,285 mil.
cz-only (tokens)	628 mil.	1,407 mil.	1,250 mil.	3,285 mil.
cz-only (words)	**526 mil.**	**1,143 mil.**	**982 mil.**	**2,652 mil.**

Table 1: Sizes of the particular parts of the corpus during the downloading and cleaning process.

smaller chunks of text, such as paragraphs, leads to the increasing number of classification errors.

As a consequence, we have developed a tool that deals with shorter texts more successfully. It is based on word n-grams estimated from the Wikipedia content. Currently, it uses word unigrams (top 100 000 most frequent words for every language) and is able to distinguish 49 languages.

Table 1 shows the final corpus size after performing unaccented-words and language filter (and leaving only correct Czech), Figure 1 shows in more detail the language composition of the data detected by our tools.

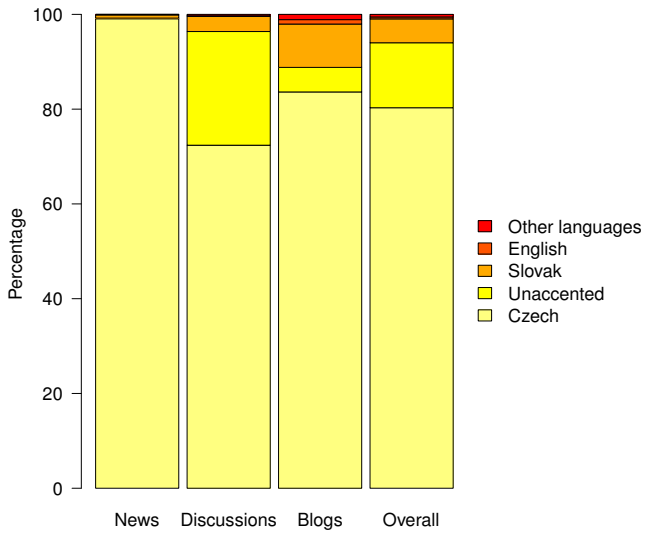

Figure 1: Results of the language filtering module.

5. Automatic linguistic processing

Our corpus is automatically linguistically processed using state-of-the-art morphological analysis (Hajič, 2004), version CZ110622a, state-of-the-art averaged perceptron POS tagger (Spoustová et al., 2009), implementation Featurama[3], feature set neopren, and a maximum spanning tree dependency parser (McDonald et al., 2005), version 0.4.3c[4]. The tagger and the parser were trained on the standard data sets from (Hajič et al., 2006).

6. Comparison with current corpora

Following our previous article (Spoustová et al., 2010), we would like to compare our new corpus to other resources available. Ideally, we would like to acquire web data that are as similar to currently available corpora as possible.

First, we focus on word and sentence measures that may be easily extracted from the texts, such as mispelled-word ratio and average sentence length. If the differences in these measures are too big, we may conclude that texts included in the web corpus differ from those in reference corpus a lot. For the comparison experiments, we chopped a 50 million token portion of the CNC SYN2005 and the three parts of our web corpus (articles, discussions, blogs). We split all the portions into 1 million-token length parts and estimate mean and standard deviation for experiments where applicable.

According to Figure 2, it turns out that in terms of average sentence length the articles part of our web corpus is quite similar to the SYN2005. This is not surprising, taking into account the SYN2005 structure (40 % fiction, 27 % technical literature, 33 % journalism).

Relatively high average sentence length in web discussions (compared to blogs) may be caused by segmentation errors (the task is difficult in some cases due to the lack of punctuation and capitalization).

The results of out-of-vocabulary words percentage measuring presented in Figure 3 are also not surprising. Texts from the articles section are written in correct Czech and most of them are professionally reviewed and proofreaded. On the contrary, in discussions and blogs "everything is allowed". We must also take into account the tolerated error-rate of the language filter and unaccented Czech filter.

In fact, the corpus comparison is quite difficult and challenging task itself. (Kilgarriff, 2001) explores several different measures of corpus similarity (and homogeneity), such as perplexity and cross-entropy of the language models, χ^2 statistics or Spearman rank correlation coefficient. Using the "Known-Similarity Corpora", he finds, that for the purpose of corpora similarity comparison, χ^2 and Spearman rank methods work significantly better than the cross-entropy based ones.

[3] http://sf.net/projects/featurama/

[4] http://sf.net/projects/mstparser/

	articles	discussions	blogs	SYN2005
articles	0.941 (0.046)	0.053	0.240	0.707
discussions		0.973 (0.011)	0.630	0.143
blogs			0.980 (0.014)	0.402
SYN2005				0.937 (0.024)

Table 2: Spearman rank correlation coefficient as a measure of homogeneity and inter-corpus similarity. Homogeneity is measured using 10 random partitions of the corpus divided into two halves and the results are average and standard deviation (in brackets).

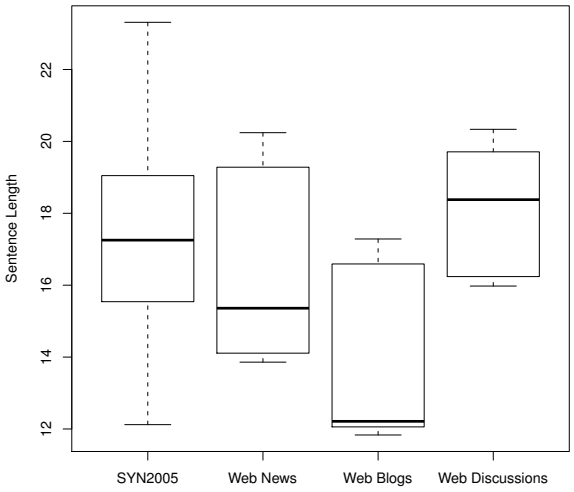

Figure 2: Average sentence length comparison of SYN2005 and particular parts of the WEB corpus.

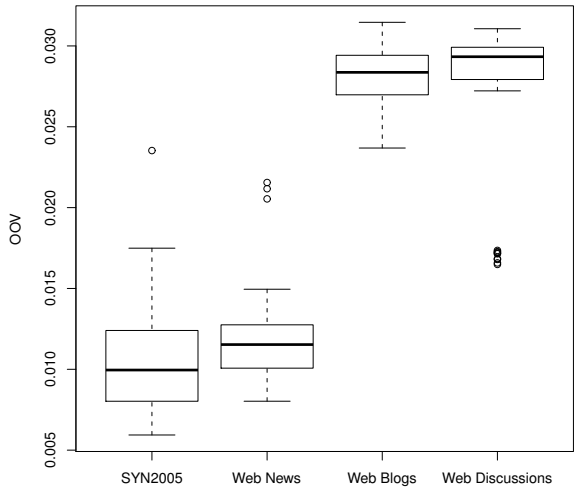

Figure 3: Box-plots of out-of-vocabulary words percentage for SYN2005 and particular parts of the WEB corpus.

For our data sets, we compute Spearman rank correlation coefficient of the distance of ranks of 500 most frequent words. The difference is small for text where common word patterns are similar. As the measure is independent of the corpora size, we can directly compare both homogeneity (intra-corpus) and similarity (inter-corpus) results.
Table 2 shows that all the (sub)corpora are quite homogenous. The highest inter-corpus similarity was achieved between web-articles and SYN2005, these corpora also have very similar homogeneity.
We can conclude that the articles sub-corpus seems to be the moct appropriate for substituting the Czech National Corpus, when necessary, while the other web sub-corpora will probably be useful for other, more specific tasks (eg. the discussions sub-corpus for dialogue systems language modelling).

7. Availability

The full version of the corpus (complete articles, blogs etc. with automatic linguistic annotation and viewable corresponding URLs) is, due to the author's law, not available for download, only for viewing and searching through our simple corpus viewer on the project website `http://hector.ms.mff.cuni.cz`
For public download, we offer following resources (also on the project's website):

- URL lists
- Shuffled sentences (annotated): articles (3.2 GB), discussions (6.1 GB), blogs (5.7 GB)
- N-gram collection (unigrams to 5grams, 2 or more occurrences, without annotation): articles, discussions, blogs, complete

The software tools (near-duplicate detection algorithm, language detection module, simple corpus viewer) are also available for download on the website `http://hector.ms.mff.cuni.cz`
As the project is finished, we cannot guarantee the availability of the Hector site in the future (in depends on financial and personal conditions of the department), but some of the resources will probably be available through the LINDAT-Clarin repository.

8. Conclusion

We have introduced new corpus of Czech web texts, which is significantly larger than Czech National Corpus, still maintaining good language quality due to a lot of human work and knowledge involved during the corpus building

process. We have also described our newly developed software tools (near-duplicate detection algorithm, language detection module), which are being released together with the data.

9. Acknowledgement

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).
The research described here was supported by the project GA405/09/0278 of the Grant Agency of the Czech Republic.

10. References

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13:422–426, July.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8-13):1157 – 1166. Sixth International World Wide Web Conference.

Moses S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, STOC '02, pages 380–388, New York, NY, USA. ACM.

CNC, 2005. *Czech National Corpus – SYN2005*. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, Prague.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank v2.0, CDROM, LDC Cat. No. LDC2006T01. Linguistic Data Consortium, Philadelphia, PA.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Miroslav Spousta, Michal Marek, and Pavel Pecina. 2008. Victor: the web-page cleaning tool. In *Proceedings of the Web as Corpus Workshop (WAC-4)*, Marrakech, Morocco.

Drahomíra "johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece, March. Association for Computational Linguistics.

Drahomíra "johanka" Spoustová, Miroslav Spousta, and Pavel Pecina. 2010. Building a web corpus of czech. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).