Typos in Czech Corpora

Marek Grác

NLP Centre, Faculty of Informatics
Masaryk University
Botanická 68a, Brno, Czech Republic
grac@fi.muni.cz

Abstract. The extended usage of written corpora not only for manual querying but also for machine learning led to the creation of massive corpora. These corpora are almost solely crawled from the internet and contain texts of various quality. Corpora that contain more typos or ungrammatical texts are more difficult to use for computational linguists and are thus a major obstacle in automatic development. In this paper we attempt to qualify some of existing Czech corpora using manually created wordlist. We will show that building such a list of frequent typos can be done without major investing when agile techniques are used.

Key words: text corpus, errors in text

1 Introduction

In the last few years the popularity of the multi-billion corpora rapidly grew. These corpora are usually built from documents that are crawled almost exclusively of internet. The language of internet differs from one which is mapped by manually compiled corpora. The main negative of corpora based on internet documents is the fact that unlike documents in manually compiled corpora (mainly newspaper, literature, ...) we are working with documents which did not pass any proofreading and some of them e.g. posts in the discussion forums are not intended to be more than a online form of communication where users do not care so much about grammar, typos and other errors. In case of English corpora we usually want to differ between documents written by a native (or qualified) speaker and those who use a mix of English and their native language [4]. For smaller languages, like Czech, we do not have to solve this problem because there are few people who actively write in Czech and are not native speakers. We still need corpora which do not contain that many various errors, for example we need to remove non-Czech documents.

For NLP applications we prefer corpora which contain only generally acceptable language without too many deviations, due to the fact that existing tools have trouble handling even “correct” language. Once we are close to solving this issue, we can try to work with the more difficult case of "real-world" language.

Based on the Czech corpora (e.g. [5], [8]), we can see that quality of corpora obtained by various methods in different times fluctuates. One of the few tests

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013, pp. 59–62, 2013. © Tribun EU 2013

done on Czech corpora is the *the-test* [7] which counts how often is token "the" used in corpora. This works because "the" is not a valid Czech word. Token "the" was used because one of the problems in crawling Czech corpora is obtaining also documents in English were "the" should occur quite frequently. Of course, even that "the" is not a valid Czech word, it should occur in large corpora, mainly in form of snippets of English texts or named entities like movies or bands. The main advantage of this text is the fact that it is very cheap to test an existing corpora. One simple query and we have results. But it tests the presence of English in corpora, what is just a part of a possibly "broken" texts.

2 Building typo language resource

In our research we would like to have a more fine-grained testing. We have decided to create a database of the most frequent Czech typos. Creating such database requires a lot of resources because manual annotation is necessary - as we do not known how to distinguish unknown word from incorrect one. We have decided to reuse techniques from agile development for creating a language resource based on proposals in [2]. The main points which we have to satisfy:

- **have an application for data:** Test quality of corpora to help us remove "broken" sentences or documents, so we can focus on problems of existing tools like morphological desambiguation, chunkers or syntactic parsers.
- **obtain a data to annotate automatically:** Data were obtained from large crawled corpus czTenTen which contains over five billion tokens. Two possible approaches can be used here. We can choose if we want to work with tokens or lemmas. If we decide to go with the lemma, then number of annotation data can be simplified, but we will have problems with unknown words which are guessed by specialized tools. This could results in creating a correct word lemma from ungrammatical token. This was the reason why we have decided to work with tokens directly. We have obtained the most frequent 100,000 tokens which are not in database of morphological analyser majka [6].
- **data have to be annotated in a simple environment:** Checking whether token is a valid one or not, should be (in most of the cases) easy task for native speaker. We have decided to not use context for given token. Annotators used an existing tool SySel [3] which allows them to confirm/deny if presented token is a valid token. Due to use of this existing resource and its simple interface the whole process of annotating data was done in 100 man-hours taken into account that each token was annotated by two different annotators.

After we have obtained a set of tokens that look like typo for annotators, we have selected a subset of these which we have agreed upon from the preselected set it was 32,766 tokens. This means that more than 66% of original data are unknown to used morphological analyser but they are still valid according to annotators. In the set of typo tokens, we can see several patterns like missing diacritics, tokens in non-latin alphabets or foreign words (mainly English ones). This data could be used for further research if we would like to fix them to correct versions what should be possible even without context for at least a portion of them.

3 Testing quality of corpora

Quality of corpora can be measured from different views but there are just a few of them which are easy to compute automatically. We have selected to use *the-score* metrics [7], a simple metric which measures contamination of the corpus by English words. The *the-score* is the rank of the word "the" in a list of tokens (originally words) sorted by frequency starting from the most frequent one. The higher values should implicate that the corpus is not polluted by English documents. We have run these tests on a selection of available corpora. The first ones DESAM [5] and CBB.blog [1] which represent small manually compiled corpora, SYN2K12 is an example of large balanced corpora and the other ones are taken from the web. At the table 1, as we have deeper user knowledge of these corpora we can generally agree with the-score with exception of CBB.blog. This corpus shows the-score which is similar to web crawled corpora but its language purity is much higher than in czTenTen12. Main reason for such high score is fact that corpus contains album and movies reviews which are often mentioned also by original (English) name.

name of corpus	the-score	absolute frequency
DESAM	9,200	12
CBB.blog	1,418	48
SYN2K12	7,897	18,847
czes2	56	530,289
czTenTen12	1,331	346,706
czTenTen12.clean	2,087	216,940

We have created a new metric based on the ratio of the known typos to all tokens in the corpus. The lower values represents that corpus is cleaner, we expect that there will be a correlation with quality of texts itself. In the table 2, results of this measurement are presented. We can see that the corpora are in similar order as in table 1 with CBB.blog exception which is ranked on the position which can be expected from manually collected data. In this also very interesting to see that typo-ratio of *clean* version of czTenTen12 is almost half of the original version with very similar size. The removed part of corpora has typo-ratio 12.24 % what makes this part of the corpus almost unusable for general usage.

name of corpus	count of typos in thousands	corpus size in millions	typo-ratio
DESAM	1.65	1.04	0.16 %
CBB.blog	2.32	0.81	0.29 %
SYN2K12	15,710	1294	1.21 %
czes2	7,053	465	1.52 %
czTenTen12	55,650	5436	1.02 %
czTenTen12.clean	28,482	5214	0.55 %

4 Conclusions and Future work

This paper introduces a new language resource of non-Czech tokens. The resources are built on corpora data and is highly reliable as each token was confirmed by two independent annotators. We have presented that the *typo-ratio* on large corpora gives similar results to *the-score* but it works better for the small corpora. As it eliminates problem of very small snippets of English named entities in Czech documents.

With the existing resource we can clean existing Czech corpora by removing documents which have bigger *typo-ratio* then is acceptable. In the future we plan to enhance this list with additional data to qualify a reason of including token into our resource.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

1. Marek Grác. Case study of bushbank concept. In *Pacific Asia Conference on Language, Information, and Computation, PACLIC*, 2011.
2. Marek Grác. *Rapid Development of Language Resources*. PhD thesis, Masaryk University, 2013.
3. Marek Grác, Adam Rambousek. Low-cost ontology development. In *GWC 2012 6th International Global Wordnet Conference*, 2012.
4. Simone Muller. *Discourse markers in native and non-native English discourse*, volume 138. John Benjamins, 2005.
5. Karel Pala, Pavel Rychlý, and Pavel Smrž. Desam – annotated corpus for Czech. In *SOFSEM'97: Theory and Practice of Informatics*. Springer, 1997.
6. Pavel Šmerk. *Towards morphological disambiguation of Czech*. PhD thesis, Ph. D. thesis proposals, Masaryk University, 2007.
7. Vít Suchomel. Recent Czech Web Corpora. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, 2012.
8. Vít Suchomel and Jan Pomikálek. Efficient web crawling for large text corpora. In *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*, 2012.