Building a corpus of evaluative sentences in multiple domains

Jana Šindlerová
Charles University in Prague
sindlerova@ufal.mff.cuni.cz

Kateřina Veselovská
Charles University in Prague
veselovska@ufal.mff.cuni.cz

1 Introduction

The area of sentiment analysis (see Liu 2009), or in other words, the automatic extraction of subjective information from a given text, has gained much attention lately, both in the commercial sphere (search for information on customer/consumer attitudes) and in public media (opinion polls vs. web sentiment analyses in politics, cf. recent presidential elections in Czech Republic). As a consequence of an increasing number of user-generated data, we witness a growing need to classify it with respect to the opinion expressed in it. However, to perform this task automatically, we first need to create and annotate a representative corpus of evaluative data and explore it from the linguistic point of view.

In this talk we describe our work on building and annotating corpora intended for the task of sentiment analysis in Czech, and developing classifiers based on this data to prove its credibility. Besides, we present the newly created Czech subjectivity lexicon (see section 3).

2 Building and evaluating the corpora

In our project we have built two plain-text corpora covering two different domains. First, a corpus of news articles was acquired, containing 560,000 words in 1661 articles. A part of this corpus (about 450 articles) has been labeled for being subjective or not. About 410 segments (mostly sentences, but also headlines and subtitles) from the subjective section, i.e. 6,686 words, have been chosen for manual annotation. On this subcorpus, the annotation scheme, based on Wiebe (2002), has been established. Manual annotation has been performed by two annotators. Interannotator agreement has been measured and the analysis of places of disagreement has been made.

During the manual annotation of the news subcorpus, several issues have arisen to be solved in order to increase the interannotator agreement. In accordance with Balahur et al. (2010), we decided not to follow reader's perspective in our future work, but instead to focus on the sentiment content of the text. Moreover, our annotating experiment resulted in a strong need for capturing additional features, such as keeping a separate category for good/bad news, elusive expressions (subjective, but non-polarized), or false polarity expressions (only seemingly subjective due to a metaphorical transfer of meaning).

The annotation scheme was then enhanced accordingly, and transferred to another acquired corpus, containing amateur movie reviews. To compare the results, we again chose 405 segments, let the same two people annotate them, measured agreement, and finally made another annotator disagreement analysis. For the sake of comparison, we also created a third corpus, containing data from a retail server with explicit prior item evaluation done by internet users (authors of the evaluative commentaries) themselves.

To verify the reliability of the corpora, we decided to train a standard unigram-based Naïve Bayes classifier together with a lexicon-based classifier, and compare their performance with the state-of-the-art results. We have also built a classifier based on the data annotated by the retail server customers to see the difference. The best performance measured by f-score was 0,89, i.e a value close to the state of the art (see Cui et al. 2006).

3 Subjectivity lexicon

In order to improve the task of sentiment analysis in Czech further, we have built a Czech subjectivity lexicon, SubLex 1.0 (Veselovská and Bojar 2012). A subjectivity lexicon is a list of domain-independent evaluative items bearing an inherent positive or negative value (see Wiebe 2004). These expressions can be used as key words in polarity detection task. The Czech subjectivity lexicon has been gained by automatic translation of a freely available English subjectivity lexicon[1], using Czech-English parallel corpus CzEng (Bojar and Žabokrtský 2006). After manual refinement it contains 4950 unique lemmas. The evaluative items collected in SubLex1.0 will be compared with the expressions obtained from the annotated data and used for an advanced classification of evaluative corpora.

4 Analysis of the data

In the course of the annotation process, we have realized that since we are dealing with a subjective text, it is necessary to give quite precise instructions about what should be annotated and what should not. For this reason, we have built a database of basic morphosyntactic patterns often used in evaluative sentences. Together with the newly established subjective lexicon SubLex, these two resources build a concise model of evaluative language in Czech

[1] http://www.cs.pitt.edu/mpqa/subj_lexicon.html

sentiment discourse.

5 Conclusion

On the basis of our research, we argue that the number and type of annotated features is dependent on the domain of the annotated text. Moreover, the dependency on text domain can even be seen in the ways the evaluation is expressed, both lexically and structurally. Either way, the presented data have laid foundations for further research into sentiment analysis in Czech. To conclude, using our manually annotated corpora, we helped to improve the task of automatic sentence-level polarity detection.

References

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B. and Belyaeva, J. 2010. *Sentiment Analysis in the News*, In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Cui, H., Mittal, V. and Datar, M. 2006. *Comparative experiments on sentiment classification for online product reviews*, In proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence.

Liu, B. 2009. "Sentiment Analysis and Subjectivity". Invited Chapter for the *Handbook of Natural Language Processing*, Second Edition. Marcel Dekker, Inc: New York.

Veselovská, K. and Bojar, O. 2012. *SubLex, the Czech subjectivity lexicon,* version 1.0.

Wiebe, J. 2002. *Instructions for Annotating Opinions in Newspaper Articles*, Department of Computer Science Technical Report TR-02-101 , University of Pittsburgh, Pittsburgh, PA.

Wiebe, J., T. Wilson, R. Bruce, M. Bell and Martin, M. 2004. Learning subjective language. *Computational Linguistics, 30*, 3.