Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification

Kateřina Veselovská

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Abstract. This paper introduces Czech subjectivity lexicon – the new lexical resource for sentiment analysis in Czech. The lexicon is a dictionary of 4947 evaluative items annotated with part of speech and tagged with positive or negative polarity. We describe the method for building the basic vocabulary and the criteria for its manual refinement. Also, we suggest possible enrichment of the fundamental lexicon. We evaluate the current version of the dictionary by implementing it to the classifiers for automatic polarity detection and compare the results of both plain and supplemented system.

1 Introduction

The main goal of sentiment analysis is the detection of a positive or negative polarity, or neutrality of a sentence (or, more broadly, a text). Most often this takes place by detecting the polarity items, i.e. words or phrases inherently bearing a positive or negative value. These words (phrases) can be found by training probabilistic models on manually annotated data. However, it seems profitable for classification to employ a set of the most frequent domain-independent polarity indicators as well. The polarity items are usually collected in the so-called subjectivity lexicons, i.e. corpora of lexical items carrying an intrinsic positive or negative meaning. The implementation of polarity items from the subjectivity lexicon into the data is the first step towards sentiment analysis.

2 Related Work

The issue of building a subjectivity lexicon is described e.g. in (Taboada et al., 2011) or more specifically in (Banea, Mihalcea and Wiebe, 2008). Here the authors use a small set of subjectivity words and a bootstrapping method of finding new candidates on the basis of a similarity measure. The authors get to the number of 4000 top frequent entries for the final lexicon. Other method for gaining a subjectivity lexicon – translation of an existing foreign language subjectivity lexicon – is described in (Banea, Mihalcea, Wiebe and Hassan, 2008). Mostly, the authors use subjectivity lexicons and sentiment analysis in general for machine translation purposes. They are interested in how the information about polarity should be transferred from one language to another, if the polarity can differ in the corresponding text spans and if it is possible to compile a subjectivity lexicon for the target language during the translation.

There is a number of papers dealing with the topic of building the subjectivity lexicons for particular languages (see e.g. Baklival et al., 2012, De Smedt et al., 2012, Jijkoun and Hofmann, 2009 or Peres-Rosas et al., 2012). But to our knowledge, in spite of the fact that there exists an ongoing research on sentiment analysis in Czech language (see Veselovská, 2012 or Habernal et al., 2013), there is no publicly known

subjectivity lexicon available for Czech which would help to improve the task and to reach the state-of-the-art results.

3 Czech Subjectivity Lexicon

The core of the Czech subjectivity lexicon has been gained by automatic translation of a freely available English subjectivity lexicon downloaded from http://www.cs.pitt.edu/mpqa/subj_lexicon.html. This lexicon is a part of the OpinionFinder, the system for subjectivity detection in English. The clues in this lexicon were collected from a number of both manually and automatically identified sources (see Riloff and Wiebe, 2003). For translating the data into Czech, we used parallel corpus CzEng 1.0 (Bojar and Žabokrtský, 2006) containing 15 million parallel sentences (233 million English and 206 million Czech tokens) from seven different types of sources automatically annotated at surface and deep layers of syntactic representation.

By this method, we gained 7228 evaluative expressions. However, some of the items or the assigned polarities appeared rather controversial. For this reason, the lexicon has been manually refined by an experienced annotator. After excluding the clearly non-evaluative items, the lexicon has been manually checked again for other incorrect entries. Below we mention the most significant types of inappropriate entries, revealed in the checking phase by an experienced annotator.

The most common problem was including items that are evaluative only in a rare or infrequent meaning or in a specific semantic context whereas mostly they represent non-evaluative expressions (e.g. *bouda* is in most cases used as a word for a "shed", though it can also mean "dirty trick"). The main criterion for marking the given item as evaluative was its universal usability in a broader context. Thus we excluded most of the domain-dependent items. The non-evaluativeness of the item was sometimes caused by wrong translation of the original English expression. In case they had not been presented in the lexicon yet, the correct translations were added manually.

On the other hand, we found a lot of items with twofold polarity. These were mostly intensifiers like *neuvěřitelně* ('incredibly'), quantifiers like *moc* ('too'), general modifiers or words which are frequently connected both to positive and negative meaning (like *[dobré/špatné] svědomí* – '[clear/guilty] conscience'). The different polarities should be distinguished later on by recording such words in the lexicon together with their prototypical collocations. Other instances also fall under this category of dual polarity, such as ambiguous words which can be used both in positive and negative meaning – e.g. *využít někoho*, meaning 'to abuse somebody' (negative), and *využít příležitosti*, 'to take the opportunity' (positive). We put these expressions aside for further research of their semantic features and corpus analysis of their collocations, since they seem to be crucial for more fine-grained sentiment analysis (see also Benamara et al., 2007).

A particular problem appeared to be words with an incorrect polarity value assigned. These could be divided into several categories. One of them are e.g. diminutives marked with positive polarity although they are very often used in negative (mostly ironic) sense – e.g. *svatoušek* – 'goody-goody'. Another large group consists of incorrect translations of negated words like *nečestný* – 'not honest', *nemilosrdný* – 'not forgiving' etc. In this case, the system did not take into account the negative particle preceding the given word and assigned positive polarity to all of them.

In the end we gained the final set of 4947 evaluative expressions. The most frequent items in the final set were nouns (e.g. *hulvát* – 'a boor', 1958) followed by verbs (e.g. *mít rád* – 'to like', 1699), adjectives (e.g. špatný – 'bad', 821) and adverbs (e.g. *dobře* – 'rightly/well/correctly', 469).

4 Data Sets

To test the credibility of the lexicon, we used several datasets on which we had previously trained the original classifiers (see Veselovská et al., 2012). Firstly, we worked with the data obtained from the Home section of the Czech news website Aktualne.cz (*http://aktualne.centrum.cz/*) manually identified as evaluative. At the beginning, there were approximately 560,000 words in 1661 articles, which have been categorized according to their subjectivity. Than we identified 175 articles (89,932 words) bearing some subjective information, 188 articles (45,395 words) with no polarity, and we labelled 90 articles (77,918 words) as "undecided". There still remain 1,208 articles which have not been classified yet. The annotators annotated 410 segments of texts (6,868 words, 1,935 unique lemmas). These segments were gained from 12 randomly chosen articles. Secondly, we used the data from Czech movie database, CSFD.cz (*http://www.csfd.cz/*). The data contained 405 evaluative segments annotated on polarity. Moreover, as both sets of the manually annotated data were pretty small, we also used auxiliary data, namely domestic appliance reviews from the Mall.cz (*http://www.mall.cz/*) retail server. We have worked with 10,177 domestic appliance reviews (158,955 words, 13,473 lemmas) from the Mall.cz retail server. These reviews were divided into positive (6,365) and negative (3,812) by their authors.

5 Testing the Lexicon

In our sentiment analysis experiments, we use the Naive Bayes classifier, a discriminative model which makes strong independence assumptions about its features, as minutely described in (Veselovská et al., 2012) with best results for the Mall.cz data. To test the subjectivity lexicon performance, we added two new features to the classifier, saying how many of the evaluative items of which polarity the given segment contained. So far, we have seen some slight improvement in identifying evaluative sentences on Aktualne.cz data when employing the 10-fold cross-validation (see tables 1 and 2).

Test result average:	**precision**	**recall**	**f-score**
POS	0.39	0.36	0.33
NEUTRAL	0.92	0.86	0.89
BOTH	0.0	0.0	0.0
NEG	0.61	0.71	0.65
average	0.84	0.81	0.82

Table 1. Aktualne.cz without subjectivity lexicon

Test result average:	**precision**	**recall**	**f-score**
POS	0.24	0.33	0.26
NEUTRAL	0.93	0.85	0.89
BOTH	0.0	0.0	0.0
NEG	0.64	0.74	0.68
average	0.85	0.81	0.83

Table 2. Aktualne.cz with subjectivity lexicon

On the other hand, we have not seen any significant improvement either on the Mall.cz or on CSFD.cz data (see tables 3, 4, 5 and 6) so far.

Test result average:	precision	recall	f-score
POS	0.93	0.92	0.92
NEG	0.85	0.88	0.87
average	0.90	0.90	0.90

Table 3. Mall.cz without subjectivity lexicon

Test result average:	precision	recall	f-score
POS	0.93	0.92	0.92
NEG	0.86	0.88	0.87
average	0.90	0.90	0.90

Table 4. Mall. cz with subjectivity lexicon

Test result average:	precision	recall	f-score
POS	0.66	0.79	0.71
NEUTRAL	0.70	0.57	0.63
NEG	0.62	0.62	0.61
average	0.67	0.65	0.65

Table 5. CSFD.cz without subjectivity lexicon

Test result average:	precision	recall	f-score
POS	0.63	0.80	0.70
NEUTRAL	0.68	0.53	0.60
NEG	0.63	0.62	0.62
average	0.66	0.64	0.64

Table 6. CSFD.cz with subjectivity lexicon

The low performance might be caused by the very small size of the data and its domain-specificity (statistically, the classifier did not reach many hits in any of the data sets). As for the future, it could be useful to test the lexicon on much bigger evaluative data.

6 Future Work

In order to improve the automatic polarity classification, it would also be advantageous to enhance the subjectivity lexicon by several methods. Firstly, we could use the dictionary-based approach as described by Hu and Liu (2004) or Kim and Hovy (2004) and grow the basic set of words by searching for their synonyms in Czech WordNet (Pala and Ševeček, 1999).

Secondly, we could employ the corpus-based approach based on syntactic or co-occurence patterns as described in (Hatzivassiloglou & McKeown, 1997). Also, we can extend the lexicon manually by Czech evaluative idioms and other common evaluative phrases. Moreover, it would probably be useful to add back some special domain-dependent modules for the different areas of evaluation. Hereby we plan to verify the hypothesis that increasing the size of the corpus could further improve the classification.

7 Conclusion

We have built and tested a subjectivity lexicon for sentiment analysis in Czech. Comparing to the previous results reached in the field, we observed that the very first version of the lexicon did not help to improve the polarity classification significantly, so its refinement needs to be a subject of the further research. However, we introduced the unique Czech subjectivity lexicon which can still serve as a lexical resource e.g. for semantic analysis or evaluative language research.

8 Acknowledgement

The work on this project has been supported by the GAUK 3537/2011 grant and by SVV project number 267 314. 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).

References

Bakliwal, A., Piyush, A. and V. Varma (2012). Hindi Subjective Lexicon: A Lexical Resource for Hindi Adjective Polarity Classification. In Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012).

Banea, C., Mihalcea, R., Wiebe, J. and S. Hassan (2008). Multilingual subjectivity analysis using machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 127-135). Association for Computational Linguistics.

Banea, C., Mihalcea, R. and J. Wiebe (2008). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008).

Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., and V. S. Subrahmanian (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

De Smedt, T. and W. Daelemans (2012). Vreselijk mooi! (terribly beautiful): A subjectivity lexicon for dutch adjectives. In Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012).

Habernal, I., Ptáček, T. and J. Steinberger (2013). Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In Proceedings of the 4th Workshop on Computatio-nal Approaches to Subjectivity, Sentiment and So-cial Media Analysis (pp 65-74).

Hatzivassiloglou, V. and K. R. McKeown (1997). Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.

Jijkoun, V. and K. Hofmann (2009). Generating a Non-English Subjectivity Lexicon: Relations That Matter. In Proceeding of EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference.

Kim, S. and E. Hovy (2004). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics.

Pala, K. and P. Ševeček (1999). The Czech WordNet, final report. Brno: Masarykova univerzita.

Perez-Rosas, V., Banea, C. and R. Mihalcea (2012). Learning Sentiment Lexicons in Spanish. In Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012).

Riloff, E. and J. Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003.

Taboada, M., Brooks, J., Tofiloski, M., Voll, K. and M. Stede (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, 37(2), pp. 267--307.

Veselovská, K. (2012). Sentence-level senti-ment analysis in Czech. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS 2012).

Veselovská, K., Hajič Jr., J. and J. Šindlerová (2012). Creating Annotated Resources for Polarity Classification in Czech. In Proceedings of KONVENS (pp. 296-304).

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