Creating Annotated Resources for Polarity Classification in Czech

Kateřina Veselovská, Jan Hajič, Jr. and Jana Šindlerová
Institute of Formal and Applied Linguistics
Charles University
Prague, Czech Republic
{veselovska|hajicj|sindlerova}@ufal.mff.cuni.cz

Abstract

This paper presents the first steps towards reliable polarity classification based on Czech data. We describe a method for annotating Czech evaluative structures and build a standard unigram-based Naive Bayes classifier on three different types of annotated texts. Furthermore, we analyze existing results for both manual and automatic annotation, some of which are promising and close to the state-of-the-art performance, see Cui (2006).

1 Introduction

One of the main subtasks in sentiment analysis is the polarity detection, which aims to classify documents according to the overall opinion they express, see e.g. Pang et al. (2002). Usually, the first step towards polarity detection is the generation of a subjectivity lexicon, i.e. a list of words marked for polarity, see e.g. Jijkoun and Hofmann (2009). Since up to now there was a lack of annotated resources for performing sentiment analysis tasks in Czech, our primary focus was to provide corpora that could serve as a basis for polarity detection systems, and to develop and test sentiment analysis systems on it. In the current study, we describe our annotation scheme, comment on its merits and pitfalls and their relation to the linguistic issues of sentiment analysis or the specific domain of the processed data. We also attempt to suggest possible remedies as a step towards providing a reliable annotation scheme for sentiment analysis in Czech.

2 Related Work

The very first stage of the project has been described in Veselovská (2012). Closely related work on sentence-level polarity classification is done by Wiegand and Klakow (2009), who also consider different linguistic features, e.g. part-of-speech information or negation scope, while building a classifier. Our annotation guidelines were inspired by the work of Wiebe (2002). Contrary to Wiebe, we do not take into account the type of attitude and the onlyfactive attribute yet. Some work on sentiment analysis in Czech has been done so far by Steinberger et al. (2011), who detect subjectivity in Czech news articles using parallel sentiment-annotated corpora. Although the authors only used one annotator for Czech, they noticed the same problems as our annotators did while annotating the news domain. Some of them, e.g. the discrepancy between the author's intention and reader's interpretation, was experienced also by Balahur and Steinberger (2009), who decided to redefine the task of sentiment analysis in the news significantly. They prepared new, more detailed guidelines which increased the final inter-annotator agreement considerably. However, our paper primarily aims at the comparison of the three data domains, and we decided not to explore news articles for the time being. There is a number of papers dealing with polarity detection in movie reviews, mostly using Naive Bayes classier and the International Movie Database data set, with quite good results (see e.g. Pang et al. (2002)). Therefore, we decided to use Czech Movie Database and a similar method as a starting point.

3 Basic Methodology of Annotation

There are three levels at which polarity can be annotated: the expression level, the sentence (segment)[1] level and the document level. Of these, especially the first one has been widely explored, but there is also a significant number of papers dealing with the problem of sentence-level sentiment analysis, e.g. Meena and Prabhabkar (2007). In our subjectivity annotation project, we have decided to start with the sentence level plain text annotation, with the long-term intention to implement the results (namely the subjectivity lexicon items derived from the annotation) in a richly annotated treebank. The sentence level annotation enables us to explore many useful linguistic features in the analysis, which can hardly be explored at the document level, such as part-of-speech information or features derived directly from the sentence structure, as described e.g. in Wiegand (2009). On the contrary, we still need to account for the fact that classifiers trained on a bag-of-words model usually perform worse at the sentence-level than at the document level, since the total number of words within a sentence is rather small and, as a result, feature vectors encoding sentences tend to be much sparser.

In the task of sentence-level polarity classification in Czech, we distinguish three functional evaluative components that need to be identified, following Wiebe (2004):

- the source, i.e. the person or entity that expresses or experiences the private state (the writer, someone quoted in the text etc.),
- the evaluation, expressed by polar elements, i.e. words or phrases inherently bearing a positive or negative value,
- the evaluated target.

In contrast to e.g. Wilson (2008), we restrict our analysis to evaluative opinions/states only. Moreover, we take into account and mark in the annotation some further aspects concerning a fine-grained subjectivity analysis, mainly expressions bordering the area of sentiment analysis, such as good/bad news (see Section 4.2), or elusive elements (expressions bearing evaluative power, but such that we cannot describe them in terms of standard polarity values, e.g. "controversy").

The annotation practice is based on the manual tagging of appropriate text spans, and it is performed by two independent annotators (later referred to as A and B).

4 Data Sets

We have trained and tested our classifier on several data sets. The primary motivation for our research was to create a tool for detecting the way news articles influence public opinion. Therefore, we initially worked with the data from the news website Aktualne.cz. However, the analysis of such texts has proven to be a rather difficult task in terms of manual annotation, as well as automatic processing, because there was a strong tendency to avoid strongly evaluative expressions, or even any explicit evaluation. For this reason, we also decided to use review data from Czech movie database, CSFD.cz, since movie reviews have been successfully used in the area of sentiment analysis for many other languages, see e.g. Thet et al. (2009). As both sets of the manually annotated data were pretty small, we also used auxiliary data, namely domestic appliance reviews from the Mall.cz retail server.

4.1 Aktualne.cz

There are approximately 560,000 words in 1661 articles obtained from the Home section of the Czech news website Aktualne.cz. In the first phase, we manually categorized some of the articles according to their subjectivity. 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 are 1,208 articles which have not been classified yet. Most of this data is not intended for manual processing but for various unsupervised machine learning methods in potential NLP applications.

The annotators annotated 410 segments of texts (6,868 words, 1,935 unique lemmas). These segments were gained from 12 randomly chosen

[1] We use 'sentence' and 'segment' interchangeably in this article. Every sentence is a segment, but not every segment is a sentence as linguistics would have it, as there were items like news headlines or one-word exclamations in the data.

opinion articles from Aktualne.cz. The segments are mostly sentences, but they also contain headlines and subtitles. In the sequel, we refer to annotation items as segments.

At the beginning, we tried to annotate all polar states that elicit a reaction from the reader. The primary instruction for annotators was simple: Should you like or dislike an entity occurring in a segment because of what that segment says, tag the entity accordingly. This choice of annotator perspective was motivated by the desired application: if our goal is for the computer to simulate a reader and thus develop sympathies, then the training data should reflect this process. It would also enable us to bypass the issue of identifying sources and assigning some trust parameter to them. However, combined with the requirement of neutrality towards the protagonists, this choice of perspective did impede the annotators' ability to make judgements about the presence of polarity in segments. The inter-annotator agreement was a little over 0.63 by Cohen's Kappa for all polarity classes. The annotators tagged about 30% of all the segments in total.

4.1.1 Problems in Tagging

Concerning the target annotation, we experienced various problems which can be divided into several categories.

The easily resolvable ones were the problems concerning annotators' instructions and their interpretation, namely the (insufficient) clarity of the instructions (e.g. concerning the question whether the annotators should tag the preposition as a part of the target or not), misinterpretation of the annotator instructions, or misinterpretation of some linguistic properties of the text. Due to the generality of the given task the boundary between the latter two phenomena is not very clear. In general it appeared quite difficult for the annotators to abstain from their personal sympathy or antipathy for the given target, especially because the texts deal with the controversial political situation before Czech parliamentary elections in 2010.

One of the specific problems of our annotation was the fact that all of our annotators have had a linguistic background, so they might have tended to tag sentences with some presupposedly linguistically interesting polarity item, even though the polarity lay in another expression or the sentence was not subjective at all. See (1);[2]

(1) A. Vláda schválila něco jiného, než co slibovala.

B. Vláda schválila **něco jiného**, než co slibovala.

The government approved [something else]$_B$ than what it had promised.

Here the target of the negative evaluation is actually "the government".

Further problems were caused by a vague interpretation of targets in polar sentences: in evaluative structures, there are different levels on which we can determine the targets, see (3):

(2) A. Dům *byl* před sedmi lety neúspěšně dražen, nyní je v zástavě banky.

B. *Dům* byl před sedmi lety neúspěšně dražen, nyní je v zástavě banky.

Seven years ago, [the house]$_B$ [was]$_A$ unsuccessfully auctioned; now it has been pledged to a bank.

Annotator A apparently felt as negative the fact that the house had been offered in the auction, most likely because the auction was unsuccessful, whereas annotator B perceived the house itself as the evaluated entity because it failed in the auction. Here we prefer the second option, since with respect to the overall topic of the document in question, we suppose that the reader will probably evaluate the house rather than the auction.

The above problems can also be caused by seeing the subjectivity structure *source – evaluation – target* as parallel to the syntacto-semantic structure *agent – predicate – patient*. Although these structures may be parallel (and they very often are), it is not always the case.

We found many discrepancies between the local and global polarity – while a part of the sentence being evaluative, the whole sentence appears rather neutral (or even its overall polarity is oriented in the opposite direction), see (4):

[2] From here on, the tagged expression is indicated in bold if the identified polar state is oriented positively, or in italic if negative. A and B refers to the decisions of the annotators A and B. In the free translation, square brackets and lower indexing mark the annotator's decision.

(3) A. V případě jeho kandidatury na tento post by **jej** podporovalo pouze 13% dotázaných, a to z řad voličů ČSSD a KSČM.

B. V případě jeho kandidatury na tento post by **jej** podporovalo pouze 13% dotázaných, a to z řad voličů ČSSD a KSČM.

In case of his candidacy for this post, [he]$_{AB}$ would be supported only by 13% respondents, mostly supporters of ČSSD and KSČM.

4.1.2 Possible Annotation Scheme Improvements

In order to improve the annotation scheme, we found necessary to abandon the reader's perspective, and to annotate not only targets, but also sources and expressions. Originally, we hoped that taking the readers perspective could prove advantageous for the identification of those polar indicators which are most relevant for the readers. However, it turned out that it is hard to identify reader-oriented polarity (and its orientation) while keeping the sources and targets anonymous. Therefore we find more useful to separate the task of identifying subjective structures and the assignment of relevance to the reader.

Another option might be to abandon the requirement for neutrality and extend the number of annotators, ideally to a representative number of readership, e. g. by means of the so-called crowdsourcing (such as Amazon's Mechanical Turk, see also Akkaya et al. (2010), or similar large scale human annotation resource).

4.2 CSFD data

In the second phase of the project, we decided to use data more convenient for the task, namely the data from Czech movie database CSFD.cz. In comparison with the previous data set, the language of these reviews was significantly more evaluative, even though it was much more domain-dependent. To compare the results, we again chose 405 segments and let the same two people annotate them. In this case, the results were slightly better, with Cohen's Kappa 0.66. However, we again experienced some problems.

4.2.1 Problems in Tagging

Perhaps the most interesting and most disturbing issue we have encountered when annotating polarity is the annotator inconsistence and mutual disagreement in establishing the borderline between polarity target and polarity expression (evaluation). A substantial part of inter-annotator disagreement in target identification lies in different perception of the extent of polarity expression with respect to the entity evaluated. This happens especially in copular sentences, both attributive and classifying.

(4) Tom Hanks je výborný herec.
Tom Hanks is an excellent actor.

In such sentences, known also as qualification by non-genuine classification, see Mathesius (1975), annotators either tag "Tom Hanks" or "actor" or "Tom Hanks;actor"[3] as targets of the polarity expression "excellent". The three alternative solutions show three different, but equally relevant ways of polarity perception. Pragmatically, the real-world entity evaluated is Tom Hanks. Syntactically, it is the headword "actor" that is modified by the qualifying adjective "wonderful". And semantically, it is the professional abilities of T. H. as an actor which are being evaluated.

(5) Kate Winslet je špatně oblečená.
Kate Winslet is poorly dressed.

As in the previous example, the target of the negative evaluation is actually both Kate Winslet and the way she dresses herself. At the beginning we have tried to capture this problem by means of copying, i.e. we kept two separate instances of a polar state, one with "Kate Winslet" as the target and "poorly dressed" as the evaluation, the other as "dressed" as the target and "poorly" as the evaluation. Doubling the polar information though did not appear to be advantageous with respect to annotators' time expenses, moreover the annotators did not succeed in capturing each single instance of the structure in question, therefore we withdrew from such treatment in favour of the more complex variant of keeping the entity as the target and the attributed quality/ability/profession etc. in the evaluation category.

[3] We use semicolon for appending discontinuous items.

Good News, Bad News During the annotation of news articles we felt the need for a separate category capturing the so-called "good" and "bad news". It appeared useful to separate sentences involving events commonly and widely accepted as "pleasant" or "unpleasant", such as triumph, wealth, or death, injury, disease, natural disaster, political failure etc., from individual subjective statements of sentiment. Interestingly it appeared quite difficult for the annotators to identify a clear-cut borderline between subjective positive/negative opinion and good/bad news, perhaps because of generally widespread metaphorical uses of the "(un)pleasant". With movie reviews, the situation was easier. First, due to the maximally subjective character of the texts, good/bad news did not appear significantly often, were easily identifiable and did not intervene much into the annotators' decision. Nevertheless, this type of disagreement did occur, e.g. in the sentence "Bůh je krutý. God is cruel." or "Dialogy jsou nepřípadné. The dialogues are inappropriate."

Non-trivial and Domain-specific Semantics As expected, the inter-annotator agreement often fails in places where the subjectivity of the sentence is hidden and embedded in metaphorically complex expressions like

(6) Všichni herci si zapomněli mimické svaly někde doma.

All the actors have forgotten their mimic muscles at home.

(7) Slovo hrdina se pro něj opravdu nehodí.

The word "hero" does not really fit him.

Moreover, sometimes the annotated polar expression serves the polarity task only within the given semantic domain. Thus whereas expressions like "špatný herec; bad (actor)" or "špatně (oblečená); poorly (dressed)" can function universally across different text genres and topics, the expressions like "psychologicky propracované (postavy); psychologically round (characters)" or "jsou střihnuty brzo; are edited too early" take the concrete polar value according to the presupposition whether we are dealing with a movie review or not. In a different text genre they could easily aquire a different polarity value, or even they could serve as neutral, non-subjective element.

4.2.2 Enhancing the Annotation Scheme

During the annotation of CSFD data we have decided to make two improvements in the annotation scheme. First we added two more polarity values, namely NONPOS and NONNEG, for capturing more fine-grained evaluation of the type "not that good" or "not that bad" respectively.

(8) Meryl není ani krásná ani výjimečná.

Meryl is neither beautiful, nor exceptional.

NONPOS

(9) Ironický nadhled v první části vlastně nebyl tak zbytečný.

The ironic detached view in the first part wasn't actually that pointless.

NONNEG

These additional labels do not equal simple "bad" or "good" values, but neither do they refer to a neutral state. Essentially, they describe a situation where the source's evaluation goes against a presupposed evaluation of the reader's. By adding additional values we risk a slight rise in the number of points of annotator's disagreement, on the other hand we are able to capture more evaluative structures and get a more thorough picture of the evaluative information in the text.

The second, rather technical improvement was the addition of a special label TOPIC for cases where the evaluation is aimed at the overall topic of the document and there is no other co-referential item in the context to which the target label could be anchored.

(10) Skvěle obsazené, vtipné, brutální, zábavné, nápadité...

Excellently casted, witty, brutal, funny, imaginative...

As in the previous case, this label should help us capture more evaluative structures that would otherwise stay unidentified. We are aware of the fact that this label might be helpful only in domains with strong evaluative character (like product reviews), but maybe less useful in case of journalistic texts in general.

4.3 Auxiliary data – Mall.cz

We have obtained 10,177 domestic appliance reviews (158,955 words, 13,473 lemmas) from the Mall.cz retail server. These reviews are divided into positive (6,365) and negative (3,812) by their authors. We found this data much easier to work with, because they are primarily evaluative by their nature and contain no complicated syntactic or semantic structures. Unlike the data from Aktualne.cz, they also contain explicit polar expressions in a prototypical use. Furthermore, they do not need to be tagged for the gold-standard annotation.

The Mall.cz data, however, do present a different set of complications: grammatical mistakes or typing errors cause noise in the form of additional lemmas and some of the reviews are also categorized incorrectly; however, compared to the problems with news articles, these are only minor difficulties and can be easily solved.

We use the Mall.cz data to verify that the automatic annotation method presented in Sect. 5 below is sound, at least on a less demanding data set.

5 Automatic Classification Experiments

In our classification scenario, we attempt to classify individual units of annotations – *segments*. The aim of these experiments is not to build state-of-the-art sentiment analysis applications, but to evaluate whether the data coming from the annotations are actually useful, where are their limits and how to eventually change the annotation guidelines to provide higher-quality data.

5.1 Classifier description

In our experimentation, we use the Naive Bayes classifier. Naive Bayes is a discriminative model which makes strong independence assumptions about its features. These assumptions generally do not hold, so the probability estimates of Naive Bayes are often wrong, however, the *classifications* it outputs can be surprisingly good.

Let $\mathcal{C}$ denote a set of polarity classes $C_1, C_2 ... C_{|\mathcal{C}|}$. The classified unit is a segment, denoted s_j from a set of segments $\mathcal{D}$. A segment s_j is composed of n lemmas $s_{j,1}, s_{j,2} \dots s_{j,n}$. (Each lemma actually has three factors: the ”real” lemma itself, its Part of Speech and the Negation tag, see 5.2. However, for the purposes of the classifier, it is important to keep negation with the real lemma, as disposing of it would make e.g. *flattering* and *unflattering* indistinguishable.) The lexicon is then the set of all lemmas in $\mathcal{D}$ and is denoted as $\mathcal{L}$. The size of the lexicon – that is, the number of distinct lemmas in the lexicon – is M. The classification features $F_i, i = 1 \dots M$ are then the presence of the i-th lemma l_i in the classified segment.

Given that the probability of classifying a segment as belonging to C is

$$p(C|F_1, F_2, \dots F_M) \propto p(C)p(F_1, F_2, \dots F_M|C) \tag{1}$$

by the Chain Rule ($p_C p(F_1, F_2, \dots F_M|C) = p(C, F_1, F_2, \dots F_M)$) and by assuming conditional independence of features $F_1 \dots F_M$ on each other it yields the following formula:

$$\begin{aligned} p(C|F_1, F_2 \dots F_M) &\propto p(C) \prod_{i=1 \dots M} p(F_i|C) \\ &\propto \log p(C) + \sum_{i=1 \dots M} \log p(F_i|C) \end{aligned} \tag{2}$$

Maximization follows by simply $argmax$-ing over both sides. The model parameters – conditional probabilities of seeing the lemma l_i in each of the classes $C_1 \dots C_{|\mathcal{C}|}$ – are estimated as MLEs $p_T(F_i|C)$ on some training data $T = w_1 \dots w_{|T|}$ with Laplacian smoothing of strength α, computed as

$$p_T(F_i|C) = \frac{freq(i, C) + \alpha}{freq(C) + \alpha|\mathcal{C}|} \tag{3}$$

where $freq(i, C)$ is the number of times lemma l_i was seen in a segment s_j labeled as belonging to the class C.

A special *UNSEEN* lemma was also added to the model, with parameters $p(C|UNSEEN)$ estimated as the marginal probabilities $p(C)$ – the probability of something generating a polarity class should be the general probability of seeing that class anyway.

5.2 Experimental settings

The experiments were carried out across the three datasets described in 4: the small richly annotated Aktualne.cz and CSFD datasets and the larger Mall.cz data which are not annotated below segment level. Exploration of those richer features, however, has not been done extensively as of yet.

When merging the annotations, we used an "eager" approach: if one annotator has tagged a segment as polar and the other as neutral, we use the polar classification; NONPOS and NONNEG are considered NEG and POS, respectively, and segments classified as BOTH and NEG (or POS) stay as BOTH. Varying the merging procedure had practically no effect on the classification.

All data for the classifiers are tagged and lemmatized using the Morce tagger (Votrubec, 2005). We retain Part of Speech and Negation tags and discard the rest.

6 Evaluation and results

In order to judge annotation quality and usefulness, we use two distinct approaches: annotator agreement and classifier performance.

6.1 Annotator agreement

On segment level, we measured whether the annotators would agree on identifying polar segments (unlabeled agreement), polar segments and their orientation (unlabeled agreement) and whether they agree on orienting segments identified as polar (orientation agreement). Additionally, we measured text anchor overlap for sources, polar expressions and targets. We used Cohen's kappa κ and f-scores on individual polarity classes (denoted f-ntr, f-plr, etc.) for agreement and f-score for text anchor overlap. Orientation was evaluated as BOTH when an annotator found both a positively and negatively oriented polar state in one segment.

For the Aktualne.cz data, out of 437 segments, the annotators tagged:

with the following agreement:

Text anchor overlap:

Annotator	1	2
Neutral	376	358
Polar	61	79
Negative	49	62
Positive	11	16
Both	1	1

Table 1: Annotator statistics on Aktualne.cz

Agreement	κ	f-ntr	f-plr	f-neg	f-pos	f-both
Unlabeled	0.659	0.944	0.714	-	-	-
Labeled	0.649	0.944	-	0.708	0.593	0
Orientation	0.818	-	-	0.975	0.889	0

Table 2: Agreement on Aktualne.cz data

Overlap	f-score
Source	0.484
Polar expr.	0.601
Target	0.562

Table 3: Overlap, Aktualne.cz

On CSFD data, out of 405 segments, the annotators identified (numbers – except for 'neutral' – are reported for polar states, thus adding up to more than 405):

Annotator	1 (JS)	2 (KV)
Neutral segs.	171	203
Polar states	348	281
Negative	150	132
Positive	180	135
Nonneg.	10	8
Nonpos.	8	6
Bad News	22	23
ESE	15	56
Elusive	2	22
False	0	10

Table 4: Annotator statistics on CSFD data

with the following agreement (reported for segments; 'both' are such segments which have been tagged with both a positive and a negative polar state):

Agreement	κ	f-ntr	f-plr	f-neg	f-pos	f-both
Unlabeled	0.659	0.809	0.850	-	-	-
Labeled	0.638	-	0.806	0.752	0.757	0.371
Orientation	0.702	-	-	0.873	0.876	0.425

Table 5: Agreement on CSFD data

Text anchor overlap:

Overlap	f-score
Source	0.750
Polar expr.	0.580
Target	0.706

Table 6: Overlap on CSFD data

The overlap scores for the Bad News, Elusive, False and ESE labels were extremely low; however, the annotators were each consistent in their reasoning behind applying these labels. A large majority of disagreement on these labels is from mislabeling. We therefore believe that agreement can be improved simply by pointing out such examples and repeating the annotation.

Aktualne.cz generally scores better on neutral segments and worse on polar ones. The similar κ would suggest that this can be put down to chance, though, because the higher prevalence of polar segments in the CSFD data makes it easier to randomly agree on them. However, the text anchor overlap shows that as far as expression-level identification goes, the annotators were much more certain on the CSFD data in what to "blame" for polarity in a given segment.

6.2 Classifier performance

The baseline for all classifier experiments was assigning the most frequent class to all segments. For all classifier experiments, we report f-score and improvement over baseline. The reported f-score is computed as an average over f-scores of individual classes weighed by their frequencies in the true data.

20-fold cross-validation was performed, with the train/test split close to 4:1. The split was done randomly, i.e. a segment had a 0.2 chance of being put into test data. No heldout data were necessary as the Laplace smoothing parameter α was set manually to 0.005; changing it didn't significantly alter results. All data were lemmatized by the Morče tagger (Votrubec, 2005).

On the Mall.cz data:

	Accuracy	f-score
Baseline	0.630	0.286
Naive Bayes	0.827	0.781

Table 7: Classifier performance on Mall.cz data

On Aktualne.cz data, the classifier was not able to perform any different from the baseline. We hypothesised, however, that this may have been due to the massive imbalance of prior probabilities and ran the experiment again with only the first 100 neutral segments.

	Accuracy	f-score
Baseline	0.787	0.694
Naive Bayes	0.787	0.694
Baseline, 100 ntr.	0.304	0.142
NB, 100 ntr.	0.778	0.531

Table 8: Classifier performance on Aktualne.cz data

On CSFD data:

	Accuracy	f-score
Baseline	0.341	0.173
Naive Bayes	0.766	0.754

Table 9: Classifier performance on CSFD.cz data

7 Conclusion

Comparing the described attempts of annotating subjectivity, we must pinpoint one observation. The success in inter-annotator agreement is dependent on the annotated text type. Unlike newspaper articles, where opinions are presented as a superstructure over informative value, and personal likes and dislikes are restricted, CSFD reviews were written with the primary intention to express subjective opinions, likes, dislikes and evaluation. Both data sets will be available to the research community for comparison.

Possibly the most important finding of the classifier experiments is that the very simple Naive Bayes polarity classifier can be trained with decent performance (at least on the film review data) with only a very modest amount of annotated data.

The fact that annotator agreement exceeded $\kappa = 0.6$ can be, given the considerable subjectivity and difficulty of the task, considered a success.

Acknowledgments

The work on this project has been supported by the GAUK 3537/2011 grant. 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

C. Akkaya, A. Conrad, J. Wiebe, and R. Mihalcea. 2010. *Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation*, NAACL-HLT 2010 Workshop on Creating Speech and Lanugage Data With Amazon's Mechanical Turk.

A. Balahur and R. Steinberger. 2009. *Rethinking Opinion Mining in Newspaper Articles: from Theory to Practice and Back*, Proceedings of the first workshop on Opinion Mining and Sentiment Analysis (WOMSA 2009).

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

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

V. Jijkoun and K. Hofmann. 2009. *Generating a Non-English Subjectivity Lexicon: Relations That Matter*, In: Proceedings of the 12th Conference of the European Chapter of the ACL.

V. Mathesius. 1975. *A Functional Analysis of Present Day English on a General Linguistic Basis*. Walter de Gruyter.

A. Meena and T. Prabhabkar. 2007. *Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis*, In: Proceedings of ECIR.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. *Thumbs up? Sentiment Classification Using Machine Learning Techniques*, In: Proceedings of EMNLP.

J. Steinberger, P. Lenkova, M. Kabadjov, R. Steinberger, and E. van der Goot. 2011. *Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora.*, Proceedings of the 8th International Conference Recent Advances in Natural Language Processing.

T.-T. Thet, J.-Ch. Na, Ch. S.-G. Khoo, and S. Shakthikumar. 2009. *Sentiment analysis of movie reviews on discussion boards using a linguistic approach*, In: Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.

K. Veselovská. 2012. *Sentence-level sentiment analysis in Czech,* In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics.

J. Votrubec (Raab). 2005. *Morphological Tagging Based on Averaged Perceptron,* In: WDS06 Proceedings of Contributed Papers.

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

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

M. Wiegand and D. Klakow. 2009. *The Role of Knowledge-based Features in Polarity Classification at Sentence Level*, In: Proceedings of the 22nd International FLAIRS Conference (FLAIRS-2009).

T. Wilson. 2008. *Fine-Grained Subjectivity Analysis.* PhD Dissertation, Intelligent Systems Program, University of Pittsburgh.