Czech Aspect-Based Sentiment Analysis: A New Dataset and Preliminary Results

Aleš Tamchyna, Ondřej Fiala, Kateřina Veselovská

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{tamchyna,fiala,veselovska}@ufal.mff.cuni.cz

Abstract: This work focuses on aspect-based sentiment analysis, a relatively recent task in natural language processing. We present a new dataset for Czech aspect-based sentiment analysis which consists of segments from user reviews of IT products. We also describe our work in progress on the task of aspect term extraction. We believe that this area can be of interest to other workshop participants and that this paper can inspire a fruitful discussion on the topic with researchers from related fields.

1 Introduction

Sentiment analysis (or opinion mining) is a field related to natural language processing (NLP) which studies how people express emotions (or opinions, sentiments, evaluations) in language and which develops methods to automatically identify such opinions.

The most typical task of sentiment analysis is to look at some short text (a sentence, paragraph, short review) and determine its *polarity* – positive, negative or neutral.

Aspect-based sentiment analysis (ABSA) refers to discovering aspects (aspect terms, opinion targets) in text and classifying their polarity. The prototypical scenario are product reviews: we assume that products have several aspects (such as size or battery life for cellphones) and we attempt to identify users' opinions on these individual aspects.

This is a more fine-grained approach than the standard formulation of sentiment analysis where the goal would be to classify the polarity of entire sentences (or even whole reviews) without regard for internal structure.

Recently, ABSA has been gaining researchers' interest, as evidenced e.g. by the two consecutive shared tasks organized within SemEval in 2014 and 2015 [7, 6].

ABSA can be roughly divided into two subtasks: (i) identification of aspects (or aspect term extraction) in text, i.e. marking (occurrences of) words which are evaluated; (ii) polarity classification, i.e. deciding whether the opinions about the identified words are positive, negative or neutral.

In this work, we introduce a new Czech dataset of product reviews annotated for ABSA and describe a preliminary method of aspect term identification which combines a rule-based approach and machine learning.

2 Dataset of IT Product Reviews

We downloaded a number of user product reviews which are publicly available on the website of an established Czech online shop with electronic devices. Each review consists of negative and positive aspects of the product. This setting pushes the customer to rate its important characteristics.

The dataset consists of two parts: (i) random short segments and (ii) longest reviews. The difference in length is reflected also in the use of language.

The first part of this dataset contains 1000 positive and 1000 negative reviews which were selected from source data and their targets were manually tagged. These targets were either aspects of the evaluated product or some general attributes (e.g. price, ease of use). The polarity of each aspect is based on whether the user submitted the segment as negative or positive. These short reviews often contain only the aspect without any evaluative phrase.

The second part of dataset consists of the longest reviews. We chose 100 of them for each polarity. These reviews represent more usual text and they tend to keep proper sentence structure. The longest review has 7057 characters.

The whole dataset provides a consistent view of language used in the on-line environment preserving both specific word forms and language structures. There is also a large amount of domain specific slang due to the origin of the text.

Dataset part	#targets	#reviews	Avg. length
Random, positive	640	1000	34.17
Random, negative	508	1000	39.72
Longest, positive	484	100	953.35
Longest, negative	353	100	855.04

Table 1: Statistics of the annotated data.

The data was annotated by a single annotator. The basic instruction was to mark all aspects or general characteristics of the product. The span of the annotated term should be as small as possible (often a single noun). For evaluation, the span can be expanded e.g. to the immediate dependency subtree of the target. Any part of speech can be marked; e.g. both "funkčnost" ("functionality") and "funkční" ("functional") should be marked.

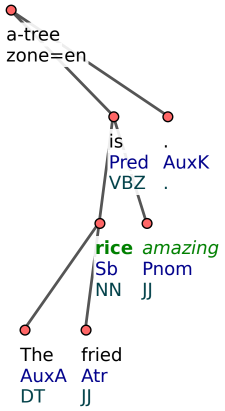

Figure 1: Dependency tree for the sentence *"The fried rice is amazing."* Morphological tags (such as *NN* for nouns) and analytical functions (e.g. *Sb* for sentence subject) are shown in the parse tree. The positive evaluative word "amazing" triggers a rule which marks "rice" as a possible aspect.

The whole dataset contains 1985 target tags; 1124 of these are positive and 861 are negative. Detailed target statistics are shown in Table 1.

The dataset is freely available for download at the following URL:

http://hdl.handle.net/11234/1-1507.

3 Pipeline

Our work is inspired by the pipeline of [15]. We run morphological analysis and tagging on the data to identify the parts of speech of words and their morphological features (e.g. case or gender for Czech). We also obtain dependency parses of the sentences. Then, we use several handcrafted rules based on syntax to mark the likely aspects in the data. Figure 1 shows a sample dependency parse tree and rule application.

Unlike [15], the core of our approach is a machine-learning model and the outputs of the rules only serve as additional "hints" (features) to help the model identify aspects.

3.1 Syntactic Rules

We use the same rules as [15], Table 2 contains their description. Here, we categorize the rules somewhat differently, their types correspond to the actual features presented to the model.

The rules are designed for *opinion target identification*, i.e. discovering targets of evaluative statements.[1] They are based on syntactic relations with evaluative words, i.e. words listed in a subjectivity lexicon for the given language.

In the example in Figure 1, the rule `vbnm_sb_adj` is triggered because "amazing" is an evaluative word and it is a predicate adjective – the word "rice", as the subject of this syntactic construction, is then marked as a likely aspect term.

Originally, the rules were written for English. Their adaptation to Czech proved very simple. We modified expressions which involved morphological tags to work with the Czech positional tagset [1]. Some of the rules included lexical items, such as the lemma *"be"* for identifying the linking verbs of predicate nominals. Simple translation of these few words to Czech sufficed in such cases.

3.2 Model

We chose linear-chain conditional random fields (CRFs) for our work [2]. In this model, aspect identification is viewed as a sequence labeling task. The input $\mathbf{x}$ are words in the sentence and the output is a labeling $\mathbf{y}$ of the same length: each word is marked as either the beginning of an aspect (B), inside an aspect (I) or outside an aspect (O).[2]

A linear-chain CRF is a statistical model. It is related to hidden Markov models (HMMs), however it is a discriminative model, not a generative one – it directly models the conditional probability of the labeling $P(\mathbf{y}|\mathbf{x})$. Linear-chain CRFs assume that the probability of the current label (B, I or O) only depends on the previous label and on the input words $\mathbf{x}$.

Formally, a linear-chain CRF is the following conditional probability distribution:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\{\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, t, \mathbf{x})\} \qquad (1)$$

Roughly speaking, $P(\mathbf{y}|\mathbf{x})$ is the score of the sentence labeling $\mathbf{y}$, exponentiated and normalized.

The score of $\mathbf{y}$ corresponds to the sum of scores for labels y_t at each position $t \in \{1, ...T\}$ in the sentence. The score at position t is the product between the values of feature functions $f_k(y_t, y_{t-1}, t, \mathbf{x})$ and their associated weights λ_k, which are estimated in the learning stage.

Feature functions can look at the current label y_t, the previous label y_{t-1} and the whole input sentence $\mathbf{x}$ (which is constant).

$Z(\mathbf{x})$ is the normalization function which sums over all possible label sequences:

$$Z(\mathbf{x}) = \sum_{\mathbf{y}'} \exp\{\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y'_t, y'_{t-1}, t, \mathbf{x})\} \qquad (2)$$

[1] The underlying assumption of this approach is that opinion targets tend to be the sought-after aspects.

[2] This "BIO" labeling scheme is common for CRFs. In practice, it brings us a consistent slight improvement as opposed to using only binary classification (inside vs. outside an aspect).

ID	Description	Example
adverb	Actor or patient of a verb with a subjective adverb.	*The pizza tastes so good.*
but_opposite	Words coordinated with an aspect with “but”.	*The food is outstanding, but everything else sucks.*
coord	Words coordinated with an aspect are also aspects.	*The excellent mussels, goat cheese and salad.*
sub_adj	Nouns modified by subjective adjectives.	*A very capable kitchen.*
subj_of_pat	Subject of a clause with a subjective patient.	*The bagel have an outstanding taste.*
verb_actant_pat	Patient of a transitive evaluative verb.	*I liked the beer selection.*
verb_actant_act	Actor of an intransitive evaluative verb.	*Their wine sucks.*
vbnm_patn	Predicative nominal (patient).	*Our favourite meal is the sausage.*
vbnm_sb_adj	Subject of predicative adjectives.	*The fried rice is amazing.*

Table 2: List of syntactic rules.

To train the model, we require training data, i.e. sentences with the labeling already assigned by a human annotator. During CRF learning, the weights λ_k are optimized to maximize the likelihood of the observed labeling in the dataset. Gradient-based optimization techniques are usually applied for learning.

At prediction time, the weights λ_k are fixed and we are looking for such a labeling $\hat{\mathbf{y}}$ which is the most probable according to the model, i.e.:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \quad (3)$$

$\hat{\mathbf{y}}$ can be found efficiently using a variant of the Viterbi algorithm (dynamic programming). In our work, we use the CRF++ toolkit[3] both for training and prediction.

3.3 Feature Set

We now describe the various feature sets evaluated in this work.

Surface features. We use the surface forms of the current word, two preceding and two following words as separate features. Additionally, we extract all (four) bigrams and (three) trigrams of surface forms from this window. We also use the CRF++ *bigram* feature template without any arguments; this simply produces the concatenation of the previous and current label (y_{t-1}, y_t).

Morpho-syntactic features. We extract unigrams, bigrams and trigrams from a limited context window (identical to the above) around the current token but instead of surface forms, we look at:

- lemma,
- morphological tag,
- analytical function.

Analytical functions are assigned by the dependency parser and their values include “Sb” for subject, “Pred” for predicate etc.

Sublex features. We mark all words in the data whose lemma is found in the subjectivity lexicon. For each token in the window of size 4 around the current token (included), we extract a feature indicating whether it was marked as subjective. We also concatenate these indicator features with the surface form of the current token.

Rule features. Finally, for each type of rule, we extract features for the current token, the preceding and the following token, indicating whether the rule marked that token. Again, these features have two versions: one standalone and one concatenated with the surface form of the current token.

4 Experiments

We analyze our data using Treex [8], a modular NLP toolkit. Sentences are first tokenized and tagged using Morphodita [12]. Then we obtain their dependency parses using the MST parser [4]. We use Czech SubLex [14] is our subjectivity lexicon both for the CRF sublex features and for the rules. The rules are implemented as blocks within the Treex platform.

4.1 Results

Table 3 shows the obtained precision (P), recall (R) and f-measure (F1) for both parts of the data set. The results in all cases were acquired using 5-fold cross-validation on the training data.

Random segments. The baseline (surface-only) features achieve the best precision but the recall is very low.

Morpho-syntactic features lower the precision by a significant margin but push recall considerably. As the review data come from the “wild”, they are quite noisy; many segments are written without punctuation, reducing the benefit of using morphological analysis, let alone dependency parsing.[4]

Often, the segments are rather short, such as “Rychlé *dodání*” (“fast *delivery*”) or “*Fotky* fakt parádní.” (“*Photos* really awesome.”). This also considerably limits the

[3] `http://taku910.github.io/crfpp/`

[4] This issue could perhaps be addressed by using a spell-checker, we leave that to future work.

Feature set	Random segments (2000)			Longest reviews (200)		
	P	R	F1	P	R	F1
surface	**85.22**	36.85	51.45	47.18	8.05	13.76
+morpho-syntactic	75.88	54.17	63.21	40.17	**23.08**	29.31
+sublex	78.19	55.09	64.64	**58.74**	18.99	28.70
+rules	76.54	**57.69**	**65.79**	51.74	21.39	**30.27**

Table 3: Precision, recall and f-measure obtained using various feature sets on the two parts of the dataset.

benefit that a parser can bring – there is a major domain mismatch both in the text topic and types of sentences between the parser's training data and this dataset, so we cannot expect parsing accuracy to be high.

Most of the improvement from adding morpho-syntactic features thus probably comes from the availability of word lemmas – this allows the CRF to learn which words are frequently marked as aspects in this domain and to generalize this information beyond their current inflected form.

Adding the information from the sentiment lexicon further improves performance, though not as much as we would expect. We could possibly further increase its impact through more careful feature engineering – so far, the features only capture whether a subjective term is present in a small linear context. For example, the lemma of the evaluative word could be included in the feature.[5]

Finally, adding the output of syntactic rules further improves the results. Due to the uncommon syntactic structure of the segments, most rules were not active very often, so the space for improvement is quite limited. Yet the results show that when the rules do trigger, their output can be a useful signal for the CRF.

The observed improvement in recall at the slight expense of precision is in line with the results of [15] where the system based on the same rules achieved high recall and rather low precision.

Long reviews. It is immediately apparent that the long reviews are a much more difficult dataset than review segments – the best f-measure achieved on the short segments is 65.79 while here it is only 30.27. This can be explained by the lower density of aspect terms compared to random review segments and a much higher sentence length – after sentence segmentation, the average sentence length is over 29 words, compared to only 6 words for the random segments.

When using only the baseline features, the recall is extremely low. Adding morpho-syntactic features has a similar effect as for the random segments – precision is lowered but recall nearly triples.

Interestingly, adding features from the subjectivity lexicon changes the picture considerably. This feature set obtains the highest precision but recall is lower compared to both +morpho-syntactic and +rules. It may be that due to the high sentence length, sublex features help identify aspects within the short window but their presence pushes the model to ignore the more distant ones. A more thorough manual evaluation would be required to confirm this.

Finally, the addition of syntactic rules leads to the highest f-measure, even though neither recall nor precision are the best. In this dataset, possibly again thanks to the length of sentences, the rules are trigged much more often than for the random segments. Rule features can therefore have a more prominent effect on the model.

5 Related Work

In terms of using rules for ABSA, our work is inspired by [15]. Such rules can also be used iteratively to expand both the aspects and evaluative terms using the double propagation algorithm [10]. Other methods of discovering opinion targets are described, inter alia, in [3, 9, 5]. Linear-chain CRFs have been applied in sentiment analysis and they are also well suited for ABSA, they were used e.g. by the winning submission by [13] to the SemEval 2014 Task 4.

For Czech, a dataset for ABSA was published by [11]. This dataset is in the domain of restaurant reviews and closely follows the methodology of [7]. Our work focuses on reviews of IT products, naturally complementing this dataset. It should further support research in this area and enable researchers to evaluate their approaches on diverse domains.

6 Conclusion

We have presented a new dataset for ABSA in the Czech language and we have described a baseline system for the subtask of aspect term extraction.

The dataset consists of segments from user reviews of IT products with the annotation of aspects and their polarity.

The system for aspect term extraction is based on linear-chain CRFs and uses a number of surface and linguistically-informed features. On top of these features, we have shown that task-specific syntactic rules can provide useful input to the model.

Utility of the syntactic rules could be further evaluated on other domains (such as the Czech restaurant reviews) or languages (e.g. using the official SemEval data sets)

[5] CRF++ feature templates do not offer a simple way to achieve this without also generating a large number of uninformative feature types.

and the impact of individual rules could be thoroughly analyzed across these data sets.

Acknowledgements

This research was supported by the grant GA15-06894S of the Grant Agency of the Czech Republic and by the SVV project number 260 224. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

References

[1] Jan Hajič and Barbora Vidová-Hladká. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of the COLING - ACL Conference*, pages 483–490, 1998.

[2] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[3] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, 2005.

[5] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 171–180, New York, NY, USA, 2007. ACM.

[6] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June 2015. Association for Computational Linguistics.

[7] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[8] Martin Popel and Zdeněk Žabokrtský. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *IceTAL 2010*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304. Iceland Centre for Language Technology (ICLT), Springer, 2010.

[9] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA, 2005.

[10] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27, March 2011.

[11] Josef Steinberger, Tomáš Brychcín, and Michal Konkol. Aspect-level sentiment analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Baltimore, USA, June 2014. Association for Computational Linguistics.

[12] Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[13] Zhiqiang Toh and Wenting Wang. DLIREC: Aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 235–240, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[14] Kateřina Veselovská and Ondřej Bojar. Czech SubLex 1.0, 2013.

[15] Kateřina Veselovská and Aleš Tamchyna. ÚFAL: Using hand-crafted rules in aspect based sentiment analysis on parsed data. In *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*, pages 694–698, Dublin, Ireland, 2014. Dublin City University, Dublin City University.