Sentence-Level Polarity Detection in a Computer Corpus

K. Veselovská

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. The paper presents a preliminary research on possible relations between the syntactic structure and the polarity of a Czech sentence by means of the so-called sentiment analysis of a computer corpus. 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 process takes place by looking for the polarity items, i.e. words or phrases inherently bearing positive or negative values. These words (phrases) are collected in the subjectivity lexicons and implemented into a computer corpus. However, when using sentences as the basic units to which sentiment analysis is applied, it is always important to look at their semantic and morphological analysis, since polarity items may be influenced by their morphological context. It is expected that some syntactic (and hypersyntactic) relations are useful for the identification of sentence polarity, such as negation, discourse relations or the level of embeddedness of the polarity item in the structure. Thus, we will propose such an analysis for a convenient source of data, the richly annotated Prague Dependency Treebank.

Introduction

Sentiment analysis (often referred to as opinion mining) tasks aim for the automatic extraction of subjective information from text and determination of speaker's attitude. The issue of subjective texts recognition has been discussed in linguistic works since early 80s and 90s, but a substantial progress in the area has started only recently with the rise of the semantically defined Web 2.0 which is based on user-generated content, e.g. social networks and weblogs [see *Ruppenhofer, Somasundaran, and Wiebe, 2008*].

There are two different types of text classification in opinion mining: subjectivity detection and polarity detection. In subjectivity detection the task is to determine whether a given text represents an opinion or a fact – or more precisely whether given information is factual or nonfactual, whereas the aim of polarity detection is to find whether the opinion expressed in a text is positive or negative.

Polarity is mostly indicated by subjective elements, i.e. single words or more complex expressions containing positive or negative polarity (e.g. *nice*, *awful* etc.). These elements are not only frequent content words. As *Wiebe et al.* [2004] states it: "Purely syntactic or morphological devices may also be subjective elements". This means that polarity items are subject to influences of sentence or larger text span context (e.g. negation or changes in aspect in both Czech and English) and thus can be profitably explored in a syntactic treebank.

Sentence-Level Polarity Classification

The main goal of sentence-level polarity detection is to decide whether a given sentence expresses either an overall positive or negative opinion. Thus, all sentences to be classified are assumed to be subjective and carrying either positive or negative overall polarity. There are several reasons why to investigate polarity detection at the sentence level. It is obvious that polarity classification at the sentence level is more fine-grained than document-level polarity classification, because every word has to be interpreted correctly (e.g. in English one needs to determine whether *like* is a verb and hence a positive polar expression or just a preposition; in Czech, we need to distinguish between particular senses of semantically ambiguous adjective *hrubá* etc.). Moreover, according to *Wiegand et al.* [2010] at the document level, text classification relies very much on redundancy and there are so many cues suggesting positive polarity more likely than negative polarity. Additionally, subjectivity is usually not uniformly distributed across a document, so the frequency analysis used e.g. in text summarization is not enough without knowledge of influence of particular polar expressions at the sentence level.

The most influential syntactic (and hypersyntactic) features useful for identification of sentence polarity ones are negation, sentential modality marking, discourse relations, intersentential coreferential relations and depth of the polarity item in the tree. The embeddedness of the polar node in a tree seems to be crucial for the polarity of a given sentence. In Figure 1, we can see an example of a sentence *Unfortunately, brother did a good job.* There are two polarity items in the structure, one positive and one negative, but its overall polarity is negative. Thus, we can assume that the higher the node is, the stronger influence it has.

It is also claimed that the main predicate is more predictive towards polarity than other words or that the main clause is more relevant than subordinate clause – see *Wiegand and Klakow* [2009].

The overall contribution of implementation of polarity items into a treebank is the inspection of such linguistic features and even polarity features derived from sentence structure and its usage in supervised machine learning.

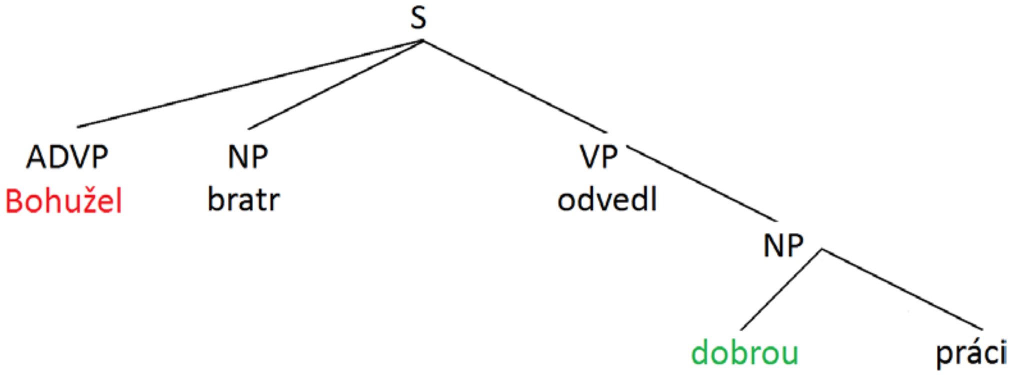

Figure 1. An illustration of the depth feature influence in the tree.

Application of Sentiment Analysis to PDT

Prague Dependency Treebank or PDT [*Hajič et al., 2006*] is a large-coverage treebank with a rich linguistic annotation (morphology, surface and deep syntax, topic-focus articulation, coreference, discourse relations etc.). Thanks to this rich annotation, it is well suited to tasks using different levels of linguistic features, like sentiment analysis. The subjective information is semantic in nature, therefore it should be embedded into the tectogrammatical layer of PDT. Despite the fact that the tectogrammatical layer seems already rather overburdened with linguistic annotation, it seems useful to keep the polarity detection at the same layer as the annotation of coreference and discourse relations, as these phenomena are closely related.

The process of application of sentiment analysis to the Prague Dependency Treebank is supposed to occur in three main phases. In the first phase of the project, it is necessary to compile a subjectivity lexicon, i.e. collection of polarity items, for Czech. The issue of building a subjectivity lexicon is concretely described e.g. in *Banea, Mihalcea and Wiebe* [2008]. The authors use a small set of subjectivity words and a bootstraping method of finding new candidates on the basis of a similarity measure. The authors get to the number of 4,000 top frequent entries for the final lexicon. The assumption is that for the purpose of PDT annotation, a sample of up to 1,000 entries should be sufficient. In case this number proves insufficient, a similar method as described in *Banea, Mihalcea and Wiebe* [2008] would be a suitable way of expanding the subjectivity lexicon further. Another method of establishing a subjectivity lexicon – translation of an existing foreign language subjectivity lexicon – is described in *Banea, Mihalcea, Wiebe and Hassan* [2008]. The authors use sentiment analysis 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 a subjectivity lexicon for the target language can be compiled during the translation. As far as Czech is concerned, the corpora can be simply based on the new Frequency Dictionary of Czech [*Čermák et al., 2004*] or the Czech Thesaurus [*Klégr, 2008*] or derived from a plain text annotation.

Secondly, the words and phrases from a subjectivity lexicon are expected to be automatically identified in the Prague Dependency Treebank and annotated using tags for “positive”, “negative”, “neutral” or “undecidable” value. The manual and automatic identification of linguistic expressions of the private states (speaker’s attitudes) is explored also in *Wilson* [2008]. Besides polarity, Wilson recognizes also intensity and attitude as important features of subjectivity expressions, with attitudes bearing two other important markers: source and target of sentiment. We believe that for the current purposes of the research on sentiment analysis in Czech, this is a too fine-grained distinction. If more tags are needed, they can be easily added during the manual control phase. The annotation should be automatic, but it will be required to make a series of manual controls of a random part of the data to ensure the reliability of annotation. Then, after tagging the data, the analysis of the annotation using statistical methods will be applied. The relationship between the number, the tag value and the position of the tagged nodes in the structure and the overall polarity of the sentence (or text) will stay in the centre of our linguistic interest. Moreover, the tagged data will thus be prepared as training data for future sentiment analysis and opinion mining experiments.

The results of the project should be applicable in many areas of Natural Language Processing, such as question answering, automatic summarization of a text, automatic dialogue systems etc.

Conclusion

Unlike some contemporary linguists [*Wiebe, Wilson, Bruce, Bell, and Martin, 2004*], we decided not to derive subjective language directly from the corpora. Though using primarily the non-contextual value of the word, the so-called prior polarity, we are aware of the possibility of context influence, therefore we include manual annotation controls into the research. We believe that the information about the amount of disagreement between a prior and contextual polarity (excluding irony) represents an important piece of information about the linguistic behaviour of subjectivity elements.

Concerning our future work, many studies [*e.g. Ruppenhofer, Somasundaran and Wiebe, 2008; Somasundaran, Namata, Wiebe and Getoor, 2009*] focus on the mutual dependency between opinion mining and discourse relations annotation. It has been pointed out that sentiment analysis is useful for the identification of discourse relations in the text, and vice versa. In this respect, our research is connected to a project aimed at the analysis and annotation of discourse relations in PDT, which is already under way.

We are not aware of any systematic research including sentiment analysis in Czech linguistic (or computational-linguistic) circles, though there are software experiments using it. Only scarcely a related study (like the one in *Smrž* [2006]) appears, but not primarily designed for Czech data. Moreover, although almost all studies of the topic mention the impact of syntactic structures, the actual research is devoted to the separate studies of individual syntactic phenomena (such as *Narayanan, Liu, and Choudhary* [2009]). Also, only a few projects use syntactically annotated corpora, although the idea is promising. Our assumption is that by using a treebank with rich linguistic annotation (including morphological, syntactical and semantic tagging, coreference, discourse and topic-focus articulation annotation) we will gain a general overview of the impact of syntactic phenomena on sentence polarity.

Acknowledgments. The present work was supported by the Charles University Grant Agency under Contract 3537/2011 and by the Czech Science Foundation under the project 201/09/H057.

References

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

Banea, C., R. Mihalcea, J. Wiebe and S. Hassan, Multilingual Subjectivity Analysis Using Machine Translation. Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), 2008.

Čermák, F. *et al.*, Frekvenční slovník češtiny. NLN, 2004.

Hajič, J., J. Panevová, E. Hajičová, P. Sgall, J. Štěpánek, J. Havelka and M. Mikulová, Prague Dependency Treebank 2.0. CD ROM. CAT: LDC2006T01, 1-58563-370-4. Linguistic Data Consortium, Univ. of Pennsylvania, Philadelphia, USA, 2006.

Klégr, A., Tezaurus jazyka českého. NLN, 2008.

Narayanan, R., B. Liu and A. Choudhary, Sentiment Analysis of Conditional Sentences. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-09). August 6-7, 2009.

Ruppenhofer, J., S. Somasundaran and J. Wiebe, Finding the Sources and Targets of Subjective Expressions. In The Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), 2008.

Smrž, P., Using WordNet for Opinion Mining. In: Proceedings of the Third International WordNet Conference, GWC 2006, Brno, CZ, MUNI, 333-335, 2006.

Somasundaran, S., G. Namata, J. Wiebe and L. Getoor, Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), 2009.

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

Wiegand, M., A. Balahur, R. Benjamin, D. Klakow and A. Montoyo, A Survey on the Role of Negation in Sentiment Analysis. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing . 60-68, 2010.

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

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