Annotation of the evaluative language in a dependency treebank

Jana Šindlerová

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic

Abstract. In the paper, we present our efforts to annotate evaluative language in the Prague Dependency Treebank 2.0. The project is a follow-up of the series of annotations of small plaintext corpora. It uses automatic identification of potentially evaluative nodes through mapping a Czech subjectivity lexicon to syntactically annotated data. These nodes are then manually checked by an annotator and either dismissed as standing in a non-evaluative context, or confirmed as evaluative. In the latter case, information about the polarity orientation, the source and target of evaluation is added by the annotator. The annotations unveiled several advantages and disadvantages of the chosen framework. The advantages involve more structured and easy-to-handle environment for the annotator, visibility of syntactic patterning of the evaluative state, effective solving of discontinuous structures or a new perspective on the influence of good/bad news. The disadvantages include little capability of treating cases with evaluation spread among more syntactically connected nodes at once, little capability of treating metaphorical expressions, or disregarding the effects of negation and intensification in the current scheme.

1 Introduction

The identification, description and analysis of evaluative language has been an important issue of computational linguistics since the rise of big data exploration. There are multiple ways to approach the issue, but basically, there are two main routes – one using the linguistically preprocessed training data to acquire reliable information about the structural properties of evaluative constructions, the other one believing in the power of unsupervised machine learning, extracting the information about evaluation from the textual data based on statistical co-occurrence of lemmata.

Within the linguistics-based approaches, a shift from plaintext annotations to the exploration of treebanks and employment of parsing mechanisms is noticeable, though both ways of data analysis have their advantages.

Plaintext annotation of evaluative states and roles is easy to learn for the annotator and in principle, does not require any specialized software. On the other hand, especially in case of large segments and less structured utterances, it may become confusing. Also, it can hardly be helped by automatic methods.

Using previously syntactically analyzed data requires availability of such data and specialized software, but it offers information helpful to the automatization of the annotation process. For example, a complex analysis of the targets as a unity of the entity and its attributes is possible, even in case of discontinuous structures. Also, it is possible to trace sources and targets of evaluation easily via anaphora resolution. In the analysis, we can make use of explicit syntactic relations, such as dependency and valency. Considering the tectogrammatic (deep syntactic) layer as the layer of capturing

evaluative relations, the problem of marking or not marking grammatical words as part of individual evaluative categories falls out of question, etc.
Our goal is to provide a sentiment annotation over an existing syntactically annotated treebank to be used in further sentiment classification and prediction tasks, and analyze its capacities to account for the persisting obstacles to the automatization of the sentiment identification and classification process. In this paper, we present an analysis of a small corpus of sentiment-annotated sentences that was created to verify the usability of the "evaluative state" annotation scheme on treebank data.

2 Related work

The current approaches to sentiment classification split basically into two branches, copying the two general approaches to machine learning: one branch promoting the use of syntactically parsed corpora as training data for the supervised learning of algorithms, and the other one favouring statistical methods (and, newly, also the neural networks) over the costly human annotation, i.e., the unsupervised learning. Both these approaches agree that using some kind of syntactic parsing yields better results than employing simple bag-of-words methods, because of the principle of compositionality of meaning, which says simply that the meaning of a compound expression is a function of the meaning of its parts and of the syntactic rules by which they are combined. Therefore, if we desire to interpret evaluation as a semantic issue in a complex and reliable way, we should use data capturing the mapping of syntactic and semantic functions.
A method to classify the sentiment polarity of a sentence based on compositional semantics was proposed, e. g., in [2]. A promising use of a treebank representation for predicting sentiment is described in [7]. The authors describe the creation of the Stanford Sentiment Treebank. The SST is an automatically parsed treebank of 11 855 movie review sentences, where each sentence was manually annotated for sentiment features by three (linguistically inexperienced) human annotators. The model trained on the SST computes sentiment using neural networks and deep learning based on the composition of meanings in the syntactic structure. The authors of [5] work with a dependency treebank and employ a probabilistic model counting polarities for each subtree. They also use a lexicon of polarity reversing words. In [6], the authors are concerned with solving metaphorical evaluations by a combination of a statistical and a rule-based system.
Though the newest studies suggest that unsupervised learning may yield optimal results at low costs in the task of automatic sentiment classification, the use of human annotated corpora lets us explore the linguistic dimension of evaluative constructions more reliably and to describe properly the evaluative patterns in everyday language.

3 Annotating evaluative language: theory and data

3.1 Plaintext annotation

The first phases of the project of capturing evaluative relations in Czech texts were carried out as series of plaintext annotations [9]. The individual parts of evaluative stance, the source, the target and the evaluative expression, were manually copied into

the cells of a spreadsheet; each evaluative stance found in the text was treated separately. Thus, e.g., the Moilanen and Pulman [4] example *The senators supporting the leader failed to praise his hopeless preventive program*, which they use for computing the overall sentiment value for the sentence, would represent (at least) three separate evaluative states, see Table 1.

The plaintext data analysis suggested there are repeating patterns for expressing evaluative meaning in the language, but did not enable a clear extraction of such patterns, due to the lack of information considering the configuration of syntactic positions of the source, target and evaluative expression in the structure.

Eval. state	Source	Evaluative expression	Target
1.	The senators	supporting	the leader
2.	The senators	failed to praise	preventive program
3.	AUTHOR	hopeless	preventive program

Table 1. Three evaluative states in the sentence *The senators supporting the leader failed to praise his hopeless preventive program*

3.2 Treebank annotation

In the second phase, we decided to use the data from the Prague Dependency Treebank 2.0 (PDT 2.0), a large and richly annotated treebank of Czech sentences [3], and apply the evaluative features to its tectogrammatic structures.

Since we need data analyzed for semantic and syntactic features, we make use of the tectogrammatical (deep syntactic) layer of PDT annotation. The choice of PDT data brought in several advantages, as well as disadvantages. It offers a complete, profound and reliable syntactic annotation with no extra annotator costs. On the other hand, a rather low amount of evaluative information is expected in the data, because the texts represent a rather objective journalist style.

The sentences of the Prague Dependency Treebank 2.0 were automatically searched for expressions matching the entries in the Czech SubLex 1.0. Czech Sublex 1.0 [8] is a Czech subjectivity lexicon, i.e., a list of subjectivity clues for sentiment analysis in Czech. It has been gained by automatic translation of a freely available English MPQA Subjectivity Lexicon [10] using a Czech-English parallel corpus CzEng 1.0 [1]. Additionally, some manual refinement of the lexicon followed in order to exclude controversial items. Finally, it contains 4626 domain-independent evaluative items (1672 positive and 2954 negative) together with their part of speech tags, polarity orientation and source English lemmas.

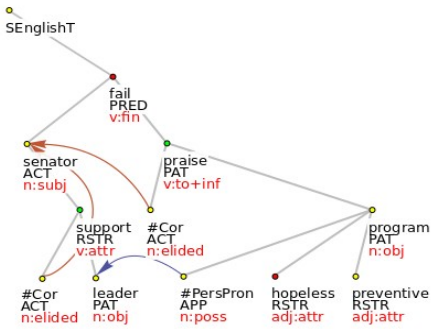

Fig. 1. Tectogrammatic representation of the sentence *The senators supporting the leader failed to praise his hopeless preventive program* with highlighted sentiments.

Fig. 1 shows a tectogrammatical representation of a typical sentence for annotation of evaluative states. There are four Sublex-suggested clues highlighted in green (for positive polarity, nodes *support* and *praise*) and red (for negative polarity, nodes *fail* and *hopeless*). The dependency links allow us to capture syntactic relations between the evaluative expressions and the sources and targets (if present overtly in the structure). The coreference arrows (blue for textual and brown for grammatical coreference) allow us to trace the lexical identity of sources and targets throughout the structure, and even beyond the sentence boundaries.

4 Annotation environment

The annotation interface was designed as an extension of the tree editor (TrEd) environment, see. Fig. 2. TrEd is a fully customizable and programmable graphical editor and viewer for tree-like structures. Among other projects, it was used as the main annotation tool for the tectogrammatical annotation of the source treebanks (PDT). It allows displaying and annotating sentential tree structures on multiple linguistic layers with a variety of tags using either the Prague Markup Language (PML) or the Treex format.

The new extension, named PML_T_Sentiment, provides a GUI supporting the entry and modification of sentiment information. The information about the part of evaluative state the individual words stand for and their possible polarity value is stored in the attribute-value matrix. The sentiment information can be changed by the annotator via use of simple macros.

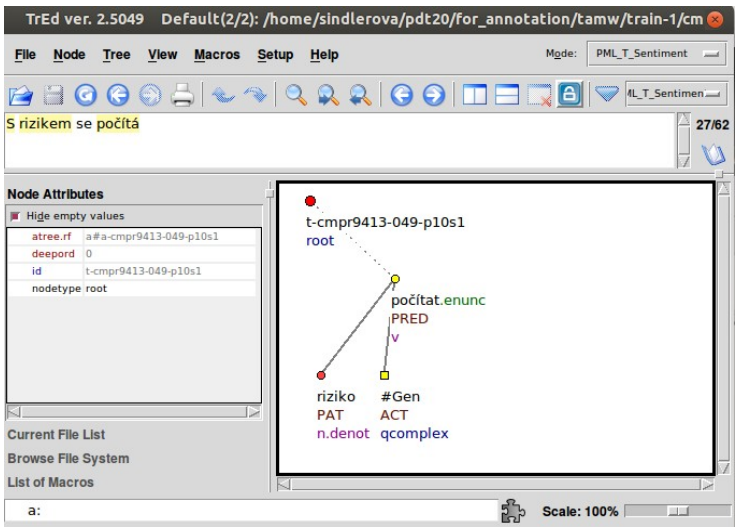

Fig. 2. Annotation environment.

5 Annotation process

The annotator is given a tectogrammatical tree for each sentence. Within the sentence, the potential candidates for evaluative nodes appear highlighted – nodes with potential positive orientation in green, nodes with potential negative orientation in red. The annotator is asked to annotate each separate highlighted node.

Annotating an evaluative node means making a decision and taking an action in each of the following issues:

1) Is the node evaluative in the given context?

An annotator is obliged to decide whether the highlighted node is in fact evaluative in the given context. If so, the annotator selects the active evaluative node, decides on its sentiment value orientation, and selects the source and target in the context. If the node is not evaluative in the given context, the sentiment highlighting and the sentiment attributes for the given node can be removed.

2) What is the source and target of the evaluative expression?

Once the node is selected, the source and target of sentiment may be annotated. If the source or target is present in the immediate sentential context, the annotator is obliged to click on it (make it active) and set it as the source or target. This inscribes the node identifier into the value of the corresponding evaluative node attribute.

If the source or target is not to be found in relatively close context, the attribute "is_extern" of the corresponding role in the attribute list of the evaluative node must be set to value "1" manually.

3) What is the polarity orientation of the evaluative node?

For each highlighted node, a polarity value is originally ascribed from the Czech SubLex by an automatic procedure. This value can be confirmed or changed manually. Immediately after the value is manually set, the node is marked as "was annotated".

Apart from the nodes suggested by the automatic comparison with Czech SubLex items, any other node in the tree may be initiated as an evaluative expression by the annotator. This is done using the function "Init Sentiment Value". By using this function, the attributes of sentiment are added into the list of attributes of the given node.

6 The pilot treebank

The "pilot sentiment treebank" contains 1044 annotated sentences of the PDT 2.0 train data section. Since our previous work showed that the interannotator agreement on evaluative state and features identification is high [9], the sentences were annotated by a single annotator only.

184 of the annotated sentences contained at least one evaluative state, positive or negative. The overall number of evaluative states found in the data is 204. This means that only 17,6 % of the sentences were evaluative.

The procedure using SubLex for identifying potential evaluative nodes highlighted 1091 candidate nodes, 754 positive and 337 negative. Strikingly, only 162 of the highlighted nodes, 79 positive and 83 negative, were confirmed as evaluative by the annotator. This means that eventually, the SubLex-based prediction does not give satisfying results. Also, the results suggest that the lexicon works far better for negative polarity clues (24,6 % predicted successfully) than for positive clues (only 10,5 % predicted successfully). We address the subjectivity lexicon limitations in the next section.

Apart from the SubLex-predicted nodes, the annotator assigned evaluation to 42 new nodes, i.e., 20,5 % of all the confirmed evaluative nodes in the pilot treebank have not been recognized by the procedure.

7 Annotation challenges

In this section, we address the common and widely known challenges to the sentiment classifying models and theories, and see how they manifest themselves in our treebank data annotation.

7.1 Subjectivity lexicon limitations

One of the underlying reasons for carrying out the annotation of sentiment in PDT was testing the feasibility of employing a subjectivity lexicon in automatic classification of structured data. While the Czech Sublex 1.0 has been originally created as a translation of the MPQA subjectivity lexicon [10], it includes lemmas falling within a much broader concept of subjectivity than the narrow concept of evaluativeness. Thus, words like *zdát se* ("to appear") or *skutečně* ("in fact"), which express (or just suggest) subjective attitude, but not specifically evaluation, appear superfluous to our purposes and are not annotated in the data.

The evaluation has been proved to be context sensitive in many cases, which makes the automatic identification of evaluative expressions even more difficult. Thus, the word *kladný* ("positive"), which comes from the subjectivity lexicon as inherently evaluative, loses its evaluative power in economic contexts (1), and other, non-evaluative words gain evaluative power in domain specific contexts (2), or when modified by an intensifier (3).

(1) I když letos a příští rok je nutné počítat se zpomalením růstu vývozu a zrychlením růstu dovozu, prognózujeme, že saldo

přesto zůstane *kladné* ve výši 300 - 600 mil. USD ročně (1 - [1,6]1.6 % HDP).
Translation: Even though it is necessary to expect a slowdown in the growth of export and a speed-up in the growth of import, we predict that the balance will remain positive*, $300-600 million a year*.
(2) Má snad mobil nějaká negativa? Ano, má. *Nepodporuje* „české“ LTE.
Translation: Does the cell-phone have any negatives? Yes, it does. It does not support *“Czech” LTE.*
(3) *Mimořádný výkon* podal Aleš Velc, který běžel druhý závod.
Translation: Aleš Velc, who ran the second race, exhibited an outstanding performance.

Unfortunately, the effects of general context on the evaluative meaning of individual words is almost as hard to be solved in treebank data, as it is in plaintext data.

7.2 Negation and other polarity reversing items

Lexical negation is usually treated as a separate grammatical node in PDT 2.0. Thus, words like *nepříznivý* (“unfavourable”) are lemmatized as positives (*příznivý,* “favourable”) and a separate “Neg” node is added as a dependant. Since the subjectivity lexicon stores negated lemmata as separate entries, this complicates the automatized matching of lexicon entries to the data. The current system matching lexicon entries to the data nodes and assigning polarity to them does not take into account polarity reversing effects of certain dependent nodes yet, therefore the automatic polarity orientation prediction usually fails with negated nodes.[1]

Apart from negation, there are other words with polarity reversing (or neutralizing) effects in the data – verbs (*znemožnit,* “prevent”), prepositions (*bez,* “without”), adverbs (*nedostatečně,* “insufficiently”, *příliš,* “excessively”). Such expressions can be stored in the form of lists, or small lexicons of polarity reversing items and (together with a set of rules for negation effects) can be employed in the system.

7.3 Bad news/good news (BGN)

Most sentiment lexicons and methodologies up to date do not discriminate evaluation from bad news/good new items properly. This is an important issue, because on one hand, BGN items in a way influence our subjective evaluative judgment of a text, on the other hand they often appear in informative, non-evaluative contexts.

The definiton of BGN was suggested in [9]. The main difference between evaluation proper and BGN lies in the fact that there is no target in case of BGN, or, more likely, the BGN items incorporate the evaluative expression and the target of evaluation both in a single word or phrase.

As the treebank data suggest, the most truth-like model will be the one showing the transition from evaluation to BGN as a scale, with unclear borderlines, since the evaluative power of BGN activates in domain specific contexts (4).

1 The same problem was experienced the other way round in the SubLex creation process. Since Sublex was translated via bilingual treebank data, wrong polarity was often assigned in contexts where negation was employed on one side of the translation, but not on the other side. These cases were then manually corrected.

(4) *Inovovali* jsme také receptury pracích prášků, *zvýšili* podíl *účinných* látek a parfémů.
Translation: We innovated *also the detergent formula, we* increased *the proportion of active ingredients and perfume.*

BGN as a phenomenon tends to follow some basic tendencies noted already in cognitive linguistics studies – we praise what is big, high, nice and healthy and we defame the opposite. The most clear example are thus the words of rising and falling (5).

(5) Ekonomika jde do vzestupu už letos.
Translation: The economics already rises this year.

7.4 Comparisons, graded sentiments

So far, the annotation scheme is only able to capture absolute polarity values. It is not designed to work with relative evaluation, which is represented linguistically, e. g., by comparison sentences, see (6).

(6) Vláda kompetence celků považuje za důležitější než jejich množství a vymezení.
Translation: The government considers the competences of the units more important than their number or delimitation.

There are two important issues connected to the treatment of comparisons in PDT data. First, the comparative degree *důležitější* ("more important") is lemmatized as *důležitý* ("important") in the treebank, and second, the second part of the comparative structure, usually elided, is represented fully in the tectogrammatic structure. The comparative word, which is usually evaluative, is thus copied in the structure (7), and therefore identified also as bearing polarity.

(7) Vláda kompetence celků považuje za důležitější než [považuje za důležité] jejich množství a vymezení.
Translation: The government considers the competences of the units more important than [it considers important] their number or delimitation.

Nevertheless, the current scheme does not take into account any scale representation of polarity strength. Therefore, the treatment of comparisons is quite difficult and fully dependent on human annotator judgement.

7.5 Metaphors

One of the almost irresolvable issues in evaluative state identification tasks is the identification of sentiment in metaphors. (8,9)

(8) Ve srovnání s vládní bitvou o počet celků z konce června *byla* tato jednání *naprostou selankou*.
Translation: Compared to the government battle over the number of units at the end of June, these negotiations were a piece of pie.

(9) Například naše zubní pasty obsadily dominantní podíl 55 procent, *čímž se nemůže pochlubit ani žádná světová firma*.
Translation: For example, our toothpastes took a dominant share of 55 %, which is something that no international company can boast of.

Since the meaning of metaphors is not derived compositionally, the treebank does not help with this task any way, nor it is easy to incorporate metaphorical expressions in the lexicon of subjective expressions due to their low frequency in language.

7.6 Complex phrases

The annotation scheme, as it is designed, ties the evaluative state to the evaluative expression matching an entry in the lexicon. From this perspective, it becomes paradoxically difficult to treat syntactically complex expressions of evaluation, as in (10). Without further improvement of the scheme, the system is not able to capture the impact of "sentiment evoking" verbs, like *považovat* ("consider").

> (10) TTI Therm *považuje* tyto návštěvy *za nejlepší způsob* dalšího zvyšování odbytu.
> *Translation: TTI Therm* considers *these visits* the best way *to increase their sales.*

8 Conclusion

We have described our efforts to annotate an existing dependency treebank with information about evaluative language. The annotations of structured data bring much light into the area of evaluative language patterns, but the currently used scheme must be further developed in order to be able to account for more complex phenomena.

1) To account for the effects of intensifiers, negation and other polarity reversing items we suggest creation of lists of polarity reversing and shifting items. Also, adding some kind of evaluation strength attribute would be beneficial.

2) The scheme should be enriched with additional attributes to account for the evaluative power of whole phrases and complex expressions, possibly also for some cases of BGN in the data.

3) It is probably not necessary to try to account for complex metaphorical expressions of evaluation.

Acknowledgments

This work has been supported by grant No. GA15-06894S of the Czech Science Foundation (GA ČR). This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth, and Sports of the Czech Republic (project LM2015071).

References

[1] Bojar, O., & Žabokrtský, Z. (2006). CzEng: Czech-English Parallel Corpus release version 0.5. *Prague Bulletin of Mathematical Linguistics*, *86*, 59–62.

[2] Choi, Y. & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: *Proceedings of the Conference on*

Empirical Methods in Natural Language Processing, pages 793–801, Association for Computational Linguistics.

[3] Hajič, Jan. (2005) Complex corpus annotation: The Prague dependency treebank. *Insight into Slovak and Czech Corpus Linguistics. Veda Bratislava* (2005): 54–73.

[4] Moilanen, K. & Pulman, S. (2007). Sentiment composition. In: *Proceedings of RANLP,* pages 378–382.

[5] Nakagawa, T., Kentaro I., & Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* Association for Computational Linguistics.

[6] Rentoumi, V., Petrakis, S., Klenner, M., Vouros, G. A., & Karkaletsis, V. (2010). United we Stand: Improving Sentiment Analysis by Joining Machine Learning and Rule Based Methods. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*.

[7] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP),* Vol. 1631, pages 1642–1653.

[8] Šindlerová, J., Veselovská, K. & Hajič, J. jr. (2014). Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon. In *Proceedings of the 16th EURALEX International Congress*, pages 405-413

[9] Veselovská, K., Hajič, J. jr. & Šindlerová, J. (2012). Creating annotated resources for polarity classification in Czech. In *Proceedings of KONVENS 2012 (PATHOS 2012 Workshop),* pages 296–304.

[10] Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing,* pages 347–354, Association for Computational Linguistics.