Originally published in: Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg. - München [u.a.]: German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019. Pp. 10-19

Distributed under a CC BY-NC-SA 4.0 license.

A Supervised Learning Approach for the Extraction of Opinion Sources and Targets from German Text

Michael Wiegand[*†], **Margarita Chikobava**[†], **Josef Ruppenhofer**[‡]
[*]Leibniz ScienceCampus, Heidelberg/Mannheim, Germany
[†]Spoken Language Systems, Saarland Informatics Campus,
Saarland University, Saarbrücken, Germany
[‡]Leibniz Institute for German Language, Mannheim, Germany
{wiegand|ruppenhofer}@ids-mannheim.de
margarita.chikobava@lsv.uni-saarland.de

Abstract

We present the first systematic supervised learning approach for the extraction of opinion sources and targets on German language data. A wide choice of different features is presented, particularly syntactic features and generalization features. We point out specific differences between opinion sources and targets. Moreover, we explain why implicit sources can be extracted even with fairly generic features. In order to ensure comparability our classifier is trained and tested on the dataset of the STEPS shared task.

1 Introduction

While there has been much research in sentiment analysis on typical text classification tasks, such as subjectivity detection, polarity classification and emotion classification, there has been notably less work on opinion role extraction. This particularly also concerns research done on languages other than English. In opinion role extraction, we distinguish between *opinion source extraction*, where the entities expressing an opinion are to be extracted, and *opinion target extraction*, where the task is to extract the entities or propositions at which sentiment is directed. For example, in (1) the sentiment expression *criticizes* has as its source *Switzerland* and as its target *North Korea*.

(1) [Switzerland $_{\text{SOURCE}}$] **criticizes** [North Korea $_{\text{TARGET}}$].
(2) [The opposition $_{\text{SOURCE}}$] **claims** [that the health service is getting fewer resources $_{\text{TARGET}}$].

In this work, we address opinion role extraction on German data. Research on this specific task and language has been kicked off by the *shared task on Source, Subjective Expression and Target Extraction from Political Speeches (STEPS)* with its two editions from 2014 (Ruppenhofer et al., 2014a) and 2016 (Ruppenhofer et al., 2016). Our experiments are carried out on these data since, to the best of our knowledge, they are the only publicly available labeled data comprising annotation for opinion role extraction on German of sufficient size from which to train a classifier. These data also allow us to directly compare our work to systems that have participated in this shared task.

In this paper, we assume that the underlying sentiment expression which evokes opinion source or opinion target has already been identified. Decoupling role extraction from the identification of sentiment expressions seems reasonable to us since previous research has focused on subjectivity detection, i.e. the detection of sentiment expressions in context. The latter task is also considerably easier in which generic and resource-poor features yield good results. Even STEPS acknowledged this by offering a subtask where sentiment expressions are already provided and thus researchers may focus solely on opinion role extraction.

The **contributions** of this paper are that we present the first in-depth study to what extent different features are relevant for the task of opinion role extraction on German data. Since we present work on German language data this means that there exist fewer NLP tools and/or tools of lesser quality. We will examine which tools actually help. While most previous approaches only focused on the extraction of either sources or targets, we consider both entity types and highlight notable differences between these tasks. We also critically assess the amount of training data that is currently available. Finally, we conduct an evaluation against previous participations in the STEPS 2016 shared task to demonstrate the effectiveness of our approach.

We acknowledge that deep learning methods have recently received considerable attention in the NLP community. However, in this work we follow a more traditional feature-based approach employing supervised learning. The reason for this is that in the area of opinion role extraction, the

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-93218

usage of deep learning methods has only produced moderate results (Katiyar and Cardie, 2016). A major caveat of deep learning methods is their reliance on distributional word representations (e.g. word embeddings). Opinion role extraction, however, is a task which relies on various types of linguistic information which are more expressive than the most robust word embeddings, such as syntactic dependency relations. Moreover, the amount of available training data for German is notably smaller than what is available for English (approximately by a factor of 10). This makes our setting fairly unfavourable for deep learning which usually outperforms traditional supervised approaches only if large amounts of labeled data are available.

2 Related Work

Like our proposed classifier, most previous approaches for opinion role extraction are supervised classifiers employing features from various information sources. They include surface-level information (Choi et al., 2005; Wiegand and Klakow, 2010), syntactic information (Choi et al., 2005; Kessler and Nicolov, 2009) and even information from semantic role labeling (Bethard et al., 2006; Kim and Hovy, 2006; Johansson and Moschitti, 2013). While particularly the latter type of information is very predictive for this task, we cannot apply it on our setting, since we are not aware of any robust semantic-role labeler for German.

Most previous research on opinion role extraction either only addressed opinion sources (Choi et al., 2005; Wiegand and Klakow, 2010; Johansson and Moschitti, 2013) or opinion targets (Kessler and Nicolov, 2009; Jakob and Gurevych, 2010). In this work, we look at both tasks. Thus we can show that there is a notable difference between these two tasks which also means that different classifier parameters and feature sets are required for those two different subtasks.

So far, work on opinion role extraction has mostly been carried out on English data. There has been some work on Chinese and Japanese as part of the NTCIR Opinion Analysis Task (Seki et al., 2007). Work on German that addresses both opinion source and target extraction has exclusively been carried out as part of the STEPS 2014 and 2016 shared tasks. There were few submissions made to the latter shared tasks. The systems presented can be divided into 3 different types:

- **rule-based approaches:** Wiegand et al. (2014) present a system that works on extraction rules defined on sentiment expressions. The system applies heavy normalization of syntactic parse trees so that simple extraction rules cover a wide range of different sentences. Wiegand et al. (2016) is an extension of that system in which further components, such as a module to detect *grammatically induced sentiment*, are added. Despite only fairly generic extraction rules, this approach produced fairly good results.
- **translation-based approaches:** Wiegand et al. (2014) also present a second system which is a supervised learning system trained on the MPQA corpus which has been automatically translated into German. That approach notably suffers from the bad translation quality.
- **supervised approaches:** Both Kriese (2016) and Wiegand et al. (2016) present a supervised classifier. While Kriese (2016) proposes models that build on *path bundles* derived from a constituency parse tree, Wiegand et al. (2016) examine an SVM trained on various features including features from syntactic parses. The results are not very conclusive as no proper feature ablation studies are carried out.

Our work substantially extends previous supervised systems as we use more features (e.g. generalization features, features derived from a constituency parse tree, subcategorization features). Moreover, we optimize various parameters and features on some development set. Thus we ensure that the features and classifiers are used in their best possible configuration. Finally, we conduct various experiments examining different feature subsets. These experiments are vital in order to make general conclusions regarding which type of information is really required for this task.

3 Data & Annotation

For our experiments we employ the labeled datasets from the STEPS 2014 shared task (Ruppenhofer et al., 2014b) and the STEPS 2016 shared task (Ruppenhofer et al., 2016) comprising 605 and 580 sentences, respectively. For STEPS 2016, the STEPS 2014 dataset was revised in order to be compatible with the new annotation scheme introduced for STEPS 2016. We use this revision of the STEPS 2014 dataset. The advantage of using datasets from the revised annotation scheme is that this scheme

Property	Freq
number of sentences	1185
average length of sentence	21
sentiment expressions	4646
sentiment expr. with neither source nor target	753
number of sources	3402
number of targets	3378
proportion of development set	10%

Table 1: Statistics of the dataset.

has been shown to produce a sufficiently high inter-annotation agreement (Ruppenhofer et al., 2016).

Since both datasets are fairly small, we merged them and conduct our experiments on the union of both datasets. Table 1 provides some descriptive statistics of our resulting dataset. 10% of the dataset were reserved as development data. On this data, we optimized various features and parameters of our classifier (§7.1).

4 Classifier and Instance Space

We pursue a supervised learning approach and decided in favor of using SVMs. As a tool, we employ SVM^{light} (Joachims, 1999). We consider the extraction of sources and targets as two completely separate tasks.

Both sources and targets always relate to a specific sentiment expression which evokes them. Therefore our instance space comprises tuples of sentiment expression and candidate opinion source phrase for sources, and sentiment expression and candidate target phrase for targets (Table 2). As a candidate source phrase, we consider all noun phrases (NPs) and preposition phrases (PPs) from the sentence in which the given sentiment expression occurs, while for targets, we consider any constituent of a sentence to be an candidate. This difference can be explained by the fact that only persons qualify as a source (hence NPs and PPs) while targets represent a more heterogeneous class of entities. For example, in (1) it is an NP representing a country while in (2) it is a complement clause representing a proposition.

5 Implicit Opinion Sources

A considerable number of opinion sources in our dataset are implicit. That is, there is no constituent in the relevant sentence that represents this opinion source. Instead the opinion source is the speaker of the utterance. For example, in (3) the sentiment expression *offensichtlich (obvious)* has no explicit source.

(3) [Die Gründe dafür TARGET] sind **offensichtlich**.
([The reasons for that TARGET] are **obvious**.)

The likelihood of an opinion source being implicit very much depends on its sentiment expression. For example, a word such as *obvious* will predominantly have an implicit source. Table 3 shows the distributions of the different source types according the part of speech of their sentiment expressions. There is clearly a correspondence. For example, of all parts of speech the likelihood of implicit sources is highest with adjectives.[1] A classifier that takes into account the part of speech of sentiment expressions is already able to make good guesses as to the presence or not of an explicit source (for example by predicting all opinion adjectives as having an implicit source and all opinion nouns having an explicit source). Further, the lexical knowledge of sentiment expressions as a feature will also be beneficial. For example, we found that more than one third of the verbal sentiment expressions having implicit sources are evoked by verbs conveying so-called *grammatically-induced sentiment* (Wiegand et al., 2016). This concerns sentiment that is conveyed by certain modalities (4)-(5).

(4) [Deshalb **müssen** wir diesen Prozess stärker ankurbeln. TARGET]
([That is why we **must** to crank this process up. TARGET])

(5) [Sie **sollten** hier ein Signal setzen. TARGET]
([You **should** send a clear message here. TARGET])

Such sentiment is evoked by frequently occurring auxiliary and modal verbs, such as *werden (will)* or *sollen (should)*. Even on comparatively small training corpora, such as ours, this information can be directly learned. That is, no manual lexicon is required for detecting such cases of sentiment as the precision of those verbs to predict an implicit source on our dataset is about 94%.

In order to enable our supervised learner to predict implicit sources, we simply need to adjust the instance space for opinion sources. In addition to explicit constituents from the sentence (see discussion above), we also add a dummy instance with an empty candidate source phrase. These instances will represent implicit sources. Indeed our exploratory experiments on the development set, as shown in Table 4, confirmed that just adding dummy instances for sources with our full feature

[1] We found that the actual proportion of implicit sources on that part of speech is actually even higher, since many sentiment adjectives having an explicit source actually turned out to be verbs erroneously tagged as adjectives.

Role	Instance	Candidate Phrases
source	<sentiment expr., candidate phrase>	all NPs, PPs and an empty dummy phrase for implicit sources (§5)
target	<sentiment expr., candidate phrase>	all phrases of a sentence

Table 2: Instances for opinion sources and targets (*the sentiment expression is always given*).

	Adj		Noun		Verb	
Source	**Freq**	**Perc**	**Freq**	**Perc**	**Freq**	**Perc**
explicit	154	27.2	1411	80.1	1164	71.1
implicit	412	72.8	350	19.9	472	28.9

Table 3: Parts of speech of implicit sources.

w/o Implicit Instances			w Implicit Instances		
Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
56.6	27.6	37.1	55.8	49.7	52.6

Table 4: Impact of implicit sources instances.

set (that includes the above features describing the part of speech and the lemma of the sentiment expression) drastically increases extraction performance for source extraction.

6 Feature Design

Our feature set is too large for us to be able to perform an evaluation on each individual feature. Instead, we group our features according to **5 meaningful dimensions** and evaluate them. In the following, we discuss those dimensions. Our complete feature set is heavily based on features employed for opinion role extraction in English. For more motivation of our feature set, we therefore refer the reader to previous work, particularly by Choi et al. (2005) and by Kessler and Nicolov (2009).

The first dimension groups our features according to the linguistic representation on which they are based. For instance, there are features that encode some semantic information, others describe syntactic or just surface-based information.

The second dimension is the focus of the feature. We distinguish between features that describe the *individual* linguistic entities involved in role extraction, that is, the sentiment expression or the candidate source/target phrase; features that describe their *relation*; and features that describe further *context* (i.e. features that focus on words other than the sentiment expression or candidate phrase).

Our third dimension divides the feature set into *simple* and *complex* features. By complex features, we understand features that require the usage of some lexical resource or a computationally intensive NLP tool (here we consider every tool more complex than a part-of-speech tagger).

The fourth dimension states whether a feature *generalizes* some lexical information or not. The generalization may be produced in a data-driven way (e.g. Brown clustering (Brown et al., 1992)) or with the help of some lexical resource (e.g. GermaNet (Hamp and Feldweg, 1997)).

The final dimension groups our feature set according to the information source it uses. By information source, we define the resource or NLP tool that is used in order to extract a particular feature. Table 5 lists all features we use and also characterizes them according to each dimension.

For part-of-speech tagging we used *TreeTagger* (Schmid, 1994), for constituency parsing the *Berkeley Parser* (Petrov et al., 2006), for dependency parsing *ParZu* (Sennrich et al., 2009), for named-entity recognition, we used the tagger by Faruqui and Padó (2010). The Brown clusters were induced with the help of *SRILM* (Stolcke, 2002). We induced 1000 clusters from the *HGC corpus*[2]. As a subcategorization lexicon, we used *IMSLex* (Fitschen, 2004).

7 Experiments

7.1 Parameter Optimization

Before we examine the different feature subsets, we need to optimize some feature and classification parameters. For these experiments, we always test a classifier on the development data. The classifiers are trained on the remaining data. We now list these optimized settings:

- Best level of generalization for GermaNet hypernyms (we do not just consider the direct hypernyms but also higher-up ancestors): for both sources and targets we consider hypernyms up to their third ancestors.
- Best cut-off value for length of part-of-speech sequences: 5 for sources; all sequences for targets.
- Best cut-off value for length of constituency paths: 5 for sources; 10 for targets.
- Best cut-off value for length of dependency relation paths: 5 for sources; 5 for targets.
- Best cost-parameter that adjusts the classifier to the imbalanced class distribution: $j = 5$ for sources; $j = 6$ for targets. (In opinion role extraction, like all entity extraction tasks, the entities to be extracted represent a

[2] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/hgc.html

	Dimensions				
Feature	**Representation**	**Focus**	**Simplicity**	**Generalizing**	**Info. Source**
head of sentiment expr.	word	individual	simple	no	lexical unit
head of candidate phrase	word	individual	simple	no	lexical unit
context as bag of words	word	context	simple	no	lexical unit
Is candidate phrase first phrase of sentence?	surface	individual	simple	no	other
orientation of candidate phrase in relation to sentiment expr.	surface	relation	simple	no	other
distance between candidate phrase and sentiment expr.	surface	relation	simple	no	other
cluster id of head of sentiment expr.	semantic	individual	complex	yes	Brown clustering
cluster id of head of candidate phrase	semantic	individual	complex	yes	Brown clustering
cluster ids of context words	semantic	relation	complex	yes	Brown clustering
named entity of candidate phrase	semantic	individual	complex	yes	named-entity tagging
synset id(s) of head of sentiment expr.	semantic	individual	complex	yes	GermaNet
synset id(s) of head of candidate phrase	semantic	individual	complex	yes	GermaNet
GermaNet word class of head of sentiment expr.	semantic	individual	complex	yes	GermaNet
GermaNet word class of head of candidate phrase	semantic	individual	complex	yes	GermaNet
GermaNet word class of words in context	semantic	relation	complex	yes	GermaNet
pos of head of sentiment expr.	syntactic	individual	simple	no	pos tagging
pos of head of candidate phrase	syntactic	individual	simple	no	pos tagging
pos sequence between candidate phrase and sentiment expr.	syntactic	relation	simple	no	pos tagging
subcategorization frame according to subcat. lexicon	syntactic	individual	complex	no	subcat. lexicon
number of arguments on subcategorization frame according to subcat. lexicon	syntactic	individual	complex	no	subcat. lexicon
phrase label of candidate phrase	syntactic	individual	complex	no	constituency parsing
tuple of phrase label of candidate phrase and pos of head of sentiment expr.	syntactic	relation	complex	no	constituency parsing
pos-tuple of head of candidate phrase and head of sentiment expr.	syntactic	relation	simple	no	constituency parsing
subcategorization frame derived from constituency tree	syntactic	individual	complex	no	constituency parsing
number of arguments in subcategorization frame derived from constituency tree	syntactic	individual	complex	no	constituency parsing
constituency label path between heads of candidate phrase and sentiment expr.	syntactic	relation	complex	no	constituency parsing
length of constituency label path between heads of candidate phrase and sentiment expr.	syntactic	relation	complex	no	constituency parsing
subcategorization frame derived from dependency tree	syntactic	individual	complex	no	dependency parsing
number of arguments on subcategorization frame derived from dependency tree	syntactic	individual	complex	no	dependency parsing
dependency relation path between heads of candidate phrase and sentiment expr.	syntactic	relation	complex	no	dependency parsing
length of dependency relation path between head of candidate phrase and sentiment expr.	syntactic	relation	complex	no	dependency parsing

Table 5: Features and their categorization along 5 dimensions.

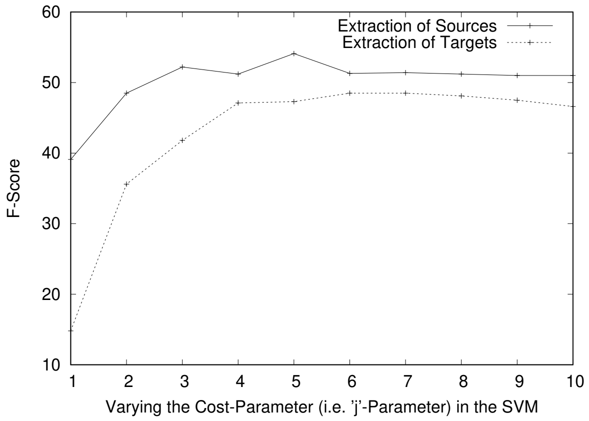

Figure 1: Optimizing the cost parameter.

minority class. This typically results in datasets with very imbalanced class distributions.)

We exemplify the importance of this optimization on the cost-parameter. Figure 1 shows the different F-scores of different cost-parameters for both source and target extraction on the development set. Clearly, the default value (i.e. $j = 1$) would only produce poor results of the classifier.

7.2 Comparison of Different Feature Groups

Given the optimal configurations determined in §7.1, we now examine the different feature groups on a 10-fold crossvalidation. We report macro-average precision, recall and F(1)-score.

Table 6 shows the performance of the individual foci and their combinations. This analysis shows that the most important focus is the set of relational features. Adding other features only yields mild increases in performance. The table also shows that regarding the other foci, there is a notable difference between the tasks. While for extraction of sources, both *individual* and *context* provide some decent F-score, on the extraction of targets they are not useful at all. While it is difficult to explain this behaviour for the context features, we found some intuitive explanation for the behaviour of the *individual* features. Opinion sources are per definition a very restricted set of entities sharing specific semantic properties. That is, only persons or groups of persons qualify as opinion sources. Therefore, a personal pronoun or the mention of a proper name (notice that our *individual* features capture this type of information), will already have a relatively high prior probability of representing a source. Targets, on the other hand, represent a much more heterogeneous group. They may be entities of various semantic types, they may even be represented by propositions (cf. (1) and (2)). This explains why targets are more dependent on relational features. That is, they can be more easily identified by their relationship towards the existing sentiment expression. For example, in both (1) and (2), the target is an object of its sentiment expression.

	Source			**Target**		
Subset	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
individual	48.2	41.2	44.4	5.4	0.2	0.3
relational	**59.5**	49.8	54.2	47.5	**40.1**	43.5
context	44.7	33.1	38.0	38.2	4.1	7.4
ind.+rel.	59.4	51.6	**55.2**	48.8	**40.1**	**44.0**
ind.+cont.	48.3	44.7	46.4	31.6	12.7	18.0
rel.+cont.	56.4	47.4	51.5	47.8	38.8	42.9
all	56.0	**54.0**	55.0	**49.1**	39.9	**44.0**

Table 6: Comparison of different foci.

	Source			**Target**		
Subset	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
all	56.0	54.0	55.0	49.1	39.9	44.0
-clustering	55.7	52.4	54.0	49.0	39.6	43.8
-GermaNet	57.0	51.6	54.1	49.9	40.6	44.7
-depend.	55.4	53.0	54.2	45.3	37.3	41.2
-constit.	55.1	49.7	52.3	46.1	35.4	40.0
-subcat	56.3	53.7	55.0	49.1	39.8	44.0
-pos	55.8	52.6	54.2	49.5	38.5	43.3
-named ent.	56.0	53.6	54.8	49.2	40.0	44.1
-other	56.1	52.3	54.1	50.6	39.1	44.1
-lexical	60.1	51.4	55.7	49.1	40.3	44.3
-dep.-const.	**52.2**	**47.6**	**49.8**	**39.6**	**32.6**	**35.8**

Table 7: Ablation experiments.

Table 7 shows some ablation experiments in which we remove one information source at a time. This gives us an indication of how unique the individual information sources are in terms of the information they contribute to the prediction of sources and targets. Only few information sources seem to carry unique information. The most notable exceptions are dependency and constituency parse information. On target extraction, we notice a notable drop in performance if either of those types of features are removed. We also removed both of these information sources at the same time to show that dependency and constituency information are not only important but are also complementary to each other.

Table 8 compares the performance of the different linguistic representations. The results are in line with the previous experiments. Word-level features are much more predictive for sources than for targets. Virtually all those features are *individual* features, so the explanation that we provided in Table 6 also applies here. Although word-level,

Subset	Source	Target
word	42.7	6.9
word+semantic	45.3	17.1
word+surface	47.2	24.4
word+semantic+surface	48.0	26.4
word+syntactic	53.6	**44.2**
word+surface+syntactic	53.6	44.1
all	**55.0**	44.0

Table 8: F-Scores of different linguistic representations.

	Source			Target		
Subset	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
simple	53.6	46.4	49.8	40.7	34.7	37.5
complex	59.9	49.7	54.3	50.3	29.9	37.5
all	56.0	54.0	55.0	49.1	39.9	44.0

Table 9: Simple and complex features.

semantic and surface features can be effectively combined, the most notable boost in performance is obtained when syntactic features are added. This is in line with our ablation experiments (Table 7) where we found that constituency and dependency parsing, in other words, *syntactic* features carry the most distinct information for this task.

Table 9 compares simple and complex features. Again, we observe notable differences between source and target extraction. While the two feature groups are on a par on target extraction, on source extraction the complex features are stronger. The combination of the feature groups is more effective on target extraction than on source extraction.

Table 10 shows the impact of generalization of both tasks. There is no clear indication that the generalization features actually help. Particularly on the extraction of targets these features are not useful at all. We explain the latter results by the fact that the generalizations are basically generalizations of the *individual* features and we pointed out in the discussion of Table 6 that those features seem to not be predictive for targets. A generalization of a completely unrelated feature is very likely to be not predictive either.

7.3 Learning Curve

The amount of labeled training data that is available to us (i.e. about 1,200 sentences) is still very small. Because of this, we computed a learning curve in order to estimate in how far increasing the amount of labeled training data would affect classification performance. Figure 2 displays a learning curve. While for sources, the curve clearly indicates that a larger amount of labeled training data is likely to increase classification performance, for targets the curve seems to be almost saturated. We already argued above that the extraction of targets is considerably more difficult than the extraction of sources. Presumably, source extraction would benefit from more labeled training data since then the classifier could get more evidence of which nouns or noun phrases are likely opinion sources and which are not. We strongly assume that due to the semantic heterogeneity of targets, such features are not effective no matter what amount of training data is available. With regard to relational/syntactic features, the current amount of labeled training data may be sufficient since there are only a handful of meaningful syntactic relationships holding between a sentiment expressions and either of its sources or targets (e.g. *subject*, *object* etc.).

	Source			Target		
Subset	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
plain	57.5	51.2	54.2	49.5	40.1	44.3
generalizat.	47.1	39.3	42.9	4.1	0.2	0.3
all	56.0	54.0	55.0	49.1	39.9	44.0

Table 10: The impact of generalization.

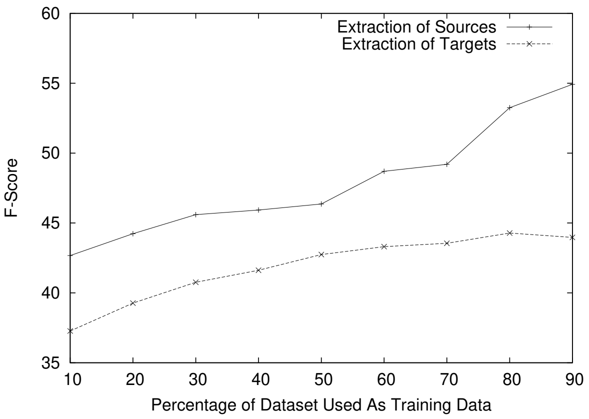

Figure 2: Learning curve on gold standard.

7.4 Comparison against Previous Classifiers

Finally, we evaluate our classifier with the full feature set against other systems that participated in the STEPS 2016 shared task. In order to produce a meaningful comparison, unlike our previous experiments, we train our classifier only on the training data from that shared task.[3] Table 11 shows the performance of the different classifiers. Overall, our proposed supervised system produces the best per-

[3] This explains why the performance of our proposed system is slightly lower than in the previous experiments.

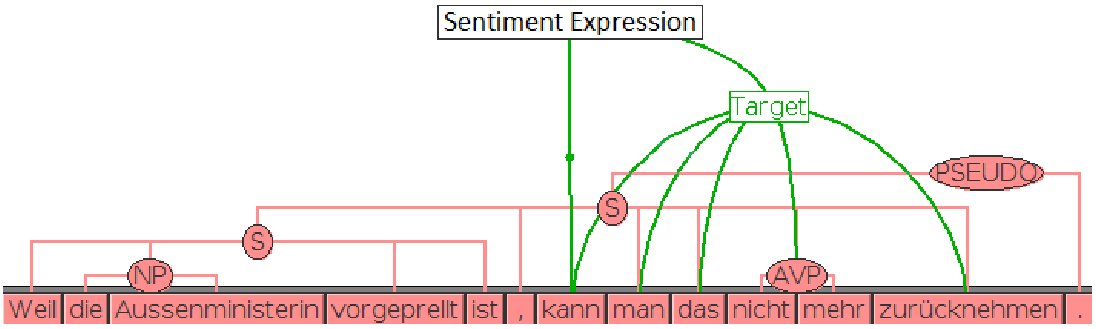

Figure 3: Illustration of opinion target covering *more* than one constituent.

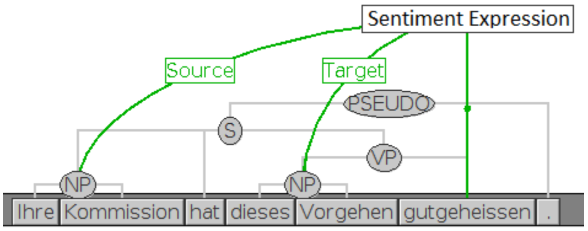

Figure 4: Illustration of opinion target covering *exactly* one constituent.

	Source		
System	**Prec**	**Rec**	**F1**
Saarland University (supervised)	59.4	38.3	46.6
Saarland University (rule based)	**59.9**	28.6	38.7
Potsdam University (supervised)	36.2	30.0	32.9
proposed system	58.0	**44.0**	**50.3**
	Target		
System	**Prec**	**Rec**	**F1**
Saarland University (supervised)	42.6	31.7	36.3
Saarland University (rule based)	**69.2**	28.9	**40.8**
Potsdam University (supervised)	37.3	22.2	27.8
proposed system	48.1	**35.0**	40.5

Table 11: Comparison with systems of the STEPS 2016 Shared Task.

formance. While on the extraction of sources, we notably outperform all other classifiers, on the extraction of targets, the rule-based system from Saarland University (Wiegand et al., 2016) is on the par with our classifier. This classifier is able to recognize instances of opinion targets that our system is unable to recognize. It concerns cases of so-called *grammatically induced sentiment* (§5). In such cases, the target typically is an entire (sub)clause. In the output of a constituency parser, these clauses often correspond to more than one constituent as illustrated in Figure 3. However, our classifier always assumes one constituent per source and target each as illustrated in Figure 4. Therefore, our approach is unable to correctly extract the above targets. In future work, we would like to combine that classifier with ours in order to hopefully obtain even a higher classification performance.

7.5 Error Analysis

Unfortunately, it is outside the scope of this paper to provide an in-depth error analysis. However, we could identify the output of syntactic parsing as a major source of error. We established in our evaluation that syntactic features are most predictive. Given that completely correct syntactic analyses on our data are rare it comes as no surprise that the overall classification performance we achieve is still comparatively low.

8 Conclusion

We presented a supervised learning approach for opinion role extraction for German. We found that there are notable differences between the extraction of opinion sources and opinion targets. Opinion targets are more difficult to handle. Even with comparably simple features, opinion sources can be extracted. For both tasks, information describing the relation between the given sentiment expression and the candidate opinion role, particularly the information drawn from syntactic parses, is most important. Generalization features do not increase classification performance much. Even though our feature set is not specifically tailored to implicit opinion sources, we are able to detect a considerable proportion. Our best classifier outperforms the best classifier which participated in the STEPS 2016 shared task. With regard to opinion target extraction, it performs on a par with the best previously reported classifier.

Acknowledgements

This research has been partially supported by the Leibniz ScienceCampus "Empirical Linguistics and Computational Modeling", funded by the Leibniz Association under grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art (MWK) of the state of Baden-Württemberg. The authors were also partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

References

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2006. Extracting Opinion Propositions and Opinion Holders using Syntactic and Lexical Cues. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 125–141. Springer-Verlag.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 355–362, Vancouver, BC, Canada.

Manaal Faruqui and Sebastian Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS*, Saarbrücken, Germany.

Arne Fitschen. 2004. *Ein computerlinguistisches Lexikon als komplexes System*. Ph.D. thesis, Institut für maschinelle Sparchverarbeitung, Universität Stuttgart.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Niklas Jakob and Iryna Gurevych. 2010. Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1045, Boston, MA, USA.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press.

Richard Johansson and Alessandro Moschitti. 2013. Relational Features in Fine-Grained Opinion Analysis. *Computational Linguistics*, 39(3):473–509.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for Joint Extraction of Opinion Entities and Relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 919–929, Berlin, Germany.

Jason S. Kessler and Nicolas Nicolov. 2009. Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, San Jose, CA, USA.

Soo-Min Kim and Eduard Hovy. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia.

Leonard Kriese. 2016. System documentation for the IGGSA Shared Task 2016. In *Proceedings of IGGSA Shared Task Workshop*, volume Bochumer Linguistische Arbeitsberichte, pages 10–13, Bochum, Germany.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 433–440, Sydney, Australia.

Josef Ruppenhofer, Roman Klinger, Julia Maria Struß, Jonathan Sonntag, and Michael Wiegand. 2014a. IGGSA Shared Tasks on German Sentiment Analysis (GESTALT). In G. Faaß and J. Ruppenhofer, editors, *Workshop Proceedings of the KONVENS Conference*, pages 164–173, Hildesheim, Germany. Universität Hildesheim.

Josef Ruppenhofer, Julia Maria Struß, Jonathan Sonntag, and Stefan Gindl. 2014b. IGGSA-STEPS: Shared Task on Source and Target Extraction from Political Speeches. *Journal for Language Technology and Computational Linguistics*, 29(1):33–46.

Josef Ruppenhofer, Julia Maria Struß, and Michael Wiegand. 2016. Overview of the IGGSA 2016 Shared Task on Source and Target Extraction from Political Speeches. In *Proceedings of IGGSA Shared Task Workshop*, Bochumer Linguistische Arbeitsberichte, pages 1–9, Bochum, Germany.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, United Kingdom.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-Hsi Chen, Noriko Kando, and Chin-Yew Lin. 2007.

Overview of Opinion Analysis Pilot Task at NTCIR-6. In *Proceedings of the NTCIR-6 Workshop Meeting*, Tokyo, Japan.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A New Hybrid Dependency Parser for German. In *Proceedings of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 115–124, Potsdam, Germany.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, Denver, CO, USA.

Michael Wiegand and Dietrich Klakow. 2010. Convolution Kernels for Opinion Holder Extraction. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 795–803, Los Angeles, CA, USA.

Michael Wiegand, Christine Bocionek, Andreas Conrad, Julia Dembowski, Jörn Giesen, Gregor Linn, and Lennart Schmeling. 2014. Saarland University's Participation in the GErman SenTiment AnaLysis shared Task (GESTALT). In G. Faaß and J. Ruppenhofer, editors, *Workshop Proceedings of the KONVENS Conference*, pages 174–184, Hildesheim, Germany. Universität Hildesheim.

Michael Wiegand, Nadisha-Marie Aliman, Tatjana Anikina, Patrick Carroll, Margarita Chikobava, Erik Hahn, Marina Haid, Katja König, Leonie Lapp, Artuur Leeuwenberg, Martin Wolf, and Maximilian Wolf. 2016. Saarland University's Participation in the Second Shared Task on Source, Subjective Expression and Target Extraction from Political Speeches (STEPS-2016). In *Proceedings of IGGSA Shared Task Workshop*, Bochumer Linguistische Arbeitsberichte, pages 14–23, Bochum, Germany.