Bochumer
Linguistische
Arbeitsberichte
18

IGGSA Shared Task on Source and Target Extraction from Political Speeches

Bochumer Linguistische Arbeitsberichte

BLA

Herausgeberin: Stefanie Dipper

Die online publizierte Reihe „Bochumer Linguistische Arbeitsberichte“ (BLA) gibt in unregelmäßigen Abständen Forschungsberichte, Abschluss- oder sonstige Arbeiten der Bochumer Linguistik heraus, die einfach und schnell der Öffentlichkeit zugänglich gemacht werden sollen. Sie können zu einem späteren Zeitpunkt an einem anderen Publikationsort erscheinen. Der thematische Schwerpunkt der Reihe liegt auf Arbeiten aus den Bereichen der Computerlinguistik, der allgemeinen und theoretischen Sprachwissenschaft und der Psycholinguistik.

The online publication series “Bochumer Linguistische Arbeitsberichte” (BLA) releases at irregular intervals research reports, theses, and various other academic works from the Bochum Linguistics Department, which are to be made easily and promptly available for the public. At a later stage, they can also be published by other publishing companies. The thematic focus of the series lies on works from the fields of computational linguistics, general and theoretical linguistics, and psycholinguistics.

Band 18 (November 2016)

Herausgeberin: Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Erscheinungsjahr	2016
ISSN	**2190-0949**

Josef Ruppenhofer, Julia Maria Struß and Michael Wiegand (Eds.)

IGGSA Shared Task on Source and Target Extraction from Political Speeches

2016

Bochumer Linguistische Arbeitsberichte

(BLA 18)

Contents

Preface

The Shared Task on Source and Target Extraction from Political Speeches (STEPS) first ran in 2014 and is organized by the Interest Group on German Sentiment Analysis (IGGSA). This volume presents the proceedings of the workshop of the second iteration of the shared task. The workshop was held at *KONVENS 2016* at Ruhr-University Bochum on September 22, 2016.

As in the first edition of the shared task the main focus of STEPS was on fine-grained sentiment analysis and offered a full task as well as two subtasks for the extraction Subjective Expressions and/or their respective Sources and Targets.

In order to make the task more accessible, the annotation schema was revised for this year's edition and an adjudicated gold standard was used for the evaluation. In contrast to the pilot task, this iteration provided training data for the participants, opening the Shared Task for systems based on machine learning approaches.

The gold standard[1] as well as the evaluation tool[2] have been made publicly available to the research community via the STEPS' website.

We would like to thank the GSCL for their financial support in annotating the 2014 test data, which were available as training data in this iteration. A special thanks also goes to Stephanie Köser for her support on preparing and carrying out the annotation of this year's test data. Finally, we would like to thank all the participants for their contributions and discussions at the workshop.

The organizers

[1] http://iggsasharedtask2016.github.io/pages/task%20description.html#data

[2] http://iggsasharedtask2016.github.io/pages/task%20description.html#scorer

Overview of the IGGSA 2016 Shared Task on Source and Target Extraction from Political Speeches

Josef Ruppenhofer*, **Julia Maria Struß**‡, **Michael Wiegand**°
* Institute for German Language, Mannheim
‡ Dept. of Information Science and Language Technology, Hildesheim University
° Spoken Language Systems, Saarland University
ruppenhofer@ids-mannheim.de
julia.struss@uni-hildesheim.de
michael.wiegand@lsv.uni-saarland.de

Abstract

We present the second iteration of IGGSA's Shared Task on Sentiment Analysis for German. It resumes the STEPS task of IGGSA's 2014 evaluation campaign: *Source, Subjective Expression and Target Extraction from Political Speeches*. As before, the task is focused on fine-grained sentiment analysis, extracting sources and targets with their associated subjective expressions from a corpus of speeches given in the Swiss parliament. The second iteration exhibits some differences, however; mainly the use of an adjudicated gold standard and the availability of training data. The shared task had 2 participants submitting 7 runs for the full task and 3 runs for each of the subtasks. We evaluate the results and compare them to the baselines provided by the previous iteration. The shared task homepage can be found at `http://iggsasharedtask2016.github.io/`.

1 Introduction

Beyond detecting the presence of opinions (or more broadly, subjectivity), opinion mining and sentiment analysis increasingly focus on determining various attributes of opinions. Among them are the polarity (or: valence) of an opinion (positive, negative or neutral), its intensity (or: strength), and also its source (or: holder) as well as its target (or: topic).

The last two attributes are the focus of the IGGSA shared task: we want to determine **whose** opinion is expressed and **what** entity or event it is about. Specific source and target extraction capabilities are required for the application of sentiment analysis to unrestricted language text, where this information cannot be obtained from meta-data and where opinions by multiple sources and about multiple, maybe related, targets appear alongside each other.

Our shared task was organized under the auspices of the Interest Group of German Sentiment Analysis[1] (IGGSA). The shared task on *Source, Subjective Expression and Target Extraction from Political Speeches (STEPS)* constitutes the second iteration of an evaluation campaign for source and target extraction on German language data. For this shared task, publicly available resources have been created, which can serve as training and test corpora for the evaluation of opinion source and target extraction in German.

2 Task Description

The task calls for the identification of subjective expressions, sources and targets in parliamentary speeches. While these texts can be expected to be opinionated, they pose the twin challenges that sources other than the speaker may be relevant and that the targets, though constrained by topic, can vary widely.

2.1 Dataset

The STEPS data set stems from the debates of the Swiss parliament (*Schweizer Bundesversammlung*). This particular data set was originally selected with the following considerations in mind. First, the source data is freely available to the public and we may re-distribute it with our annotations. We were not able to fully ascertain the copyright situation for German parliamentary speeches, which we had also considered using. Second, this type of text poses the interesting challenge of dealing with multiple sources and targets that cannot be gleaned easily from meta-data but need to be retrieved from the running text.

[1] `https://sites.google.com/site/iggsahome/`

As the Swiss parliament is a multi-lingual body, we were careful to exclude not only non-German speeches but also German speeches that constitute responses to, or comments on, speeches, heckling, and side questions in other languages. This way, our annotators did not have to label any German data whose correct understanding might rely on material in a language that they might not be able to interpret correctly.

Some potential linguistic difficulties consisted in peculiarities of Swiss German found in the data. For instance, the vocabulary of Swiss German is sometimes subtly different from standard German. For instance, the verb *vorprellen* is used in the following example rather than *vorpreschen*, which would be expected for German spoken in Germany:

(1) Es ist unglaublich: Weil die Aussenministerin vorgeprellt ist, kann man das nicht mehr zurücknehmen. (Hans Fehr, Frühjahrsession 2008, Zweite Sitzung – 04.03.2008)[2]
‘It is incredible: because the foreign secretary acted rashly, we cannot take that back again.’

In order to limit any negative impact that might come from misreadings of the Swiss German by our annotators, who were German and Austrian rather than Swiss, we selected speeches about what we deemed to be non-parochial issues. For instance, we picked texts on international affairs rather than ones about Swiss municipal governance.

The training data for the 2016 shared task comprises annotations on 605 sentences. It represents a single, adjudicated version of the three-fold annotations that served as test data in the first iteration of the shared task in 2014. The test data for the 2016 shared task was newly annotated. It consists of 581 sentences that were drawn from the same source, namely speeches from the Swiss parliament on the same set of topics as used for the training data.

Technically, the annotated STEPS data was created using the following pre-processing pipeline. Sentence segmentation and tokenization was done using OpenNLP[3], followed by lemmatization with the TreeTagger (Schmid, 1994), constituency parsing by the Berkeley parser (Petrov and Klein, 2007), and final conversion of the parse trees into TigerXML-Format using TIGER-tools (Lezius, 2002). To perform the annotation we used the Salto-Tool (Burchardt et al., 2006).[4]

2.2 Continuity with, and Differences to, Previous Annotation

Through our annotation scheme[5], we provide annotations at the expression level. No sentence or document-level annotations are manually performed or automatically derived.

As on the first iteration of the shared task, there were no restrictions imposed on annotations. The sources and targets could refer to any actor or issue as we did not focus on anything in particular. The subjective expressions could be verbs, nouns, adjectives, adverbs or multi-words.

The definition of subjective expressions (SE) that we used is broad and based on well-known prototypes. It is inspired by Wilson and Wiebe (2005)‘s use of the superordinate notion *private state*, as defined by Quirk et al. (1985): “As a result, the annotation scheme is centered on the notion of private state, a general term that covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments.”:

- evaluation (positive or negative):
  toll ‘great’, *doof* ‘stupid’
- (un)certainty:
  zweifeln ‘doubt’, *gewiss* ‘certain’
- emphasis:
  sicherlich/bestimmt ‘certainly’
- speech acts:
  sagen ‘say’, *ankündigen* ‘announce’
- mental processes:
  denken ‘think’, *glauben* ‘believe’

Beyond giving the prototypes, we did not seek to impose on our annotators any particular definition of subjective or opinion expressions from the linguistic, natural language processing or psychological literature related to subjectivity, appraisal, emotion or related notions.

[2] `http://www.parlament.ch/ab/frameset/d/n/4802/263473/d_n_4802_263473_263632.htm`

[3] `http://opennlp.apache.org/`

[4] In addition to the XML files with the subjectivity annotations, we also distributed to the shared task participants several other files containing further aligned annotations of the text. These were annotations for named entities and of dependency rather than constituency parses.

[5] See `http://iggsasharedtask2016.github.io/data/guide_2016.pdf` for the the guidelines we used.

Formally, in terms of subjective expressions, there were several noticeable changes made relative to the first iteration. First, unlike in the 2014 iteration of the shared task, punctuation marks (such as exclamation marks) could no longer be annotated. Second, while in the first iteration only the head noun of a light verb construction was identified as the subjective expression, in this iteration the light verbs were also to be included in the subjective expression. Annotators were instructed to observe the handling of candidate expressions in common dictionaries: if a light verb is mentioned as part of an entry, it should be labeled as part of the subjective expression. Thus, the combination *Angst haben* (lit. 'have fear') represents a single subjective expression, whereas in the first edition of the shared task only the noun *Angst* was treated as the subjective expression. A third change concerned compounds. We decided to no longer annotate sub-lexically. This meant that compounds such as *Staatstrauer* 'national mourning' would only be treated as subjective expressions but that we would not break up the word and label the head *-trauer* as a subjective expression and the modifier *Staats* as a Source. Instead, we label the whole word only as a subjective expression.

As before, in marking subjective expressions, the annotators were told to select minimal spans. This guidance was given because we had decided that within the scope of this shared task we would forgo any treatment of polarity and intensity. Accordingly, negation, intensifiers and attenuators and any other expressions that might affect a minimal expression's polarity or intensity could be ignored.

When labeling sources and targets, annotators were asked to first consider *syntactic and semantic dependents* of the subjective expressions. If sources and targets were locally unrealized, the annotators could annotate other phrases in the context. Where a subjective expression represented the view of the implicit speaker or text author, annotators were asked to indicate this by setting a flag *Sprecher* 'Speaker' on the the source element. Typical cases of subjective expressions are evaluative adjectives such as *toll* 'great' in (2).

(2) Das ist natürlich schon **toll**.
'Of course that's really great.'

For all three types of labels, subjective expressions, sources, and targets, annotators had the option of using an additional flag to mark an annotation as *Unsicher* 'Uncertain', if they were unsure whether the span should really be labeled with the relevant category.

In addition, instances of subjective expressions and sources could be marked as *Inferiert* 'Inferred'. In the case of subjective expressions, this covers, for instance, cases where annotators were not sure if an expression constituted a polar fact or an inherently subjective expression. In the case of sources, the 'inferred' label applies to cases where the referents cannot be annotated as local dependents but have to be found in the context. An example is sentence (3), where the source of *Strategien aufzuzeigen* 'to lay out strategies' is not a direct grammatical dependent of that complex predicate. Instead it can be found 'higher up' as a complement of the noun *Ziel* 'goal', which governs the verb phrase that *aufzuzeigen* heads.[6]

(3) Es war jedoch nicht Ziel des vorliegenden [Berichtes ^{Source}], an dieser Stelle **Strategien aufzuzeigen**, . . . zeitlichem Fokus auf das Berichtsjahr zu beschreiben.
'However, it wasn't the goal of the report at hand to lay out strategies here, . . . '

Note that, unlike in the first iteration, we decided to forego the annotation of inferred targets as we felt they would be too difficult to retrieve automatically. Also, we limited contextual annotation to the same sentence as the subjective expression. In other words, annotators could not mark source mentions in preceding sentences.

Likewise, whereas in the first iteration, the annotators were asked to use a flag *Rhetorisches Stilmittel* 'Rhetorical device' for subjective expression instances where subjectivity was conveyed through some kind of rhetorical device such as repetition, such instances were ruled out of the remit of this shared task. Accordingly, no such instances occur in our data. Even more importantly, whereas for the first iteration, we had asked annotators to also annotate polar facts and mark them with a flag, for the second iteration we decided to exclude polar facts from annotation altogether as they had led to low agreement among the annotators in the first iteration of the task. What we had called polar facts in the guidelines of the 2014 task, we would now call inferred opinions of the sort arising from events that affect their participants positively or

[6] Grammatically speaking, this is an instance of what is called control.

	2014	**2016**	
	Fleiss κ	Cohen's κ	obs.agr.
subj. expr.	0.39	0.72	0.91
sources	0.57	0.80	0.96
targets	0.46	0.60	0.80

Table 1: Comparison of IAA values for 2014 and 2016 iterations of the shared task

negatively.[7] For instance, for a sentence such as *100-year old driver crashes into school crowd*, one might infer a negative attitude of the author towards the driver, especially if the context emphasizes the driver's culpability or argues generally against letting older drivers keep their permits.

As in the first iteration, the annotation guidelines gave annotators the option to mark particular subjective expressions as *Schweizerdeutsch* 'Swiss German' when they involved language usage that they were not fully familiar with. Such cases could then be excluded or weighted differently for the purposes of system evaluation. In our annotation, these markings were in fact very rare with only one such instance in the training data and none in the test data.

2.3 Interannotator Agreement

We calculated agreement in terms of a token-based κ value. Given that in our annotation scheme, a single token can be e.g. a target of one subjective expression while itself being a subjective expression as well, we need to calculate three kappa values covering the binary distinctions between presence of each label and its absence.

In the first iteration of the shared task, we calculated a multi-κ measure for our three annotators on the basis of their annotations of the 605 sentences in the full test set of the 2014 shared task (Davies and Fleiss, 1982). For this second iteration, two annotators performed double annotation on 50 sentences as the basis for IAA calculation. For lack of resources, the rest of the data was singly annotated. We calculated Cohen's kappa values. As Table 1 suggests, inter-annotator agreement was considerably improved. This allowed participants to use the annotated and adjudicated 2014 test data as training data in this iteration of the shared task.

2.4 Subtasks

As did the first iteration, the second iteration offered a full task as well as two subtasks:

Full task Identification of subjective expressions with their respective sources and targets.

Subtask 1 Participants are given the subjective expressions and are only asked to identify opinion sources.

Subtask 2 Participants are given the subjective expressions and are only asked to identify opinion targets.

Participants could choose any combination of the tasks.

2.5 Evaluation Metrics

The runs that were submitted by the participants of the shared task were evaluated on different levels, according to the task they chose to participate in. For the full task, there was an evaluation of the subjective expressions as well as the targets and sources for subjective expressions, matching the system's annotations against those in the gold standard. For subtasks 1 and 2, we evaluated only the sources or targets, respectively, as the subjective expressions were already given.

In the first iteration of the STEPS task, we evaluated each submitted run against each of our three annotators individually rather than against a single gold-standard. The intent behind that choice was to retain the variation between the annotators. In the current, second iteration, the evaluation is simpler as we switched over to a single adjudicated reference annotation as our gold standard.

We use recall to measure the proportion of correct system annotations with respect to the gold standard annotations. Additionally, precision was calculated so as to give the fraction of correct system annotations relative to all the system annotations.

In this present iteration of the shared task, we use a strict measure for our primary evaluation of system performance, requiring precise span overlap for a match. [8]

[7] The terminology for these cases is somewhat in flux. Deng et al. (2013) talk about benefactive/malefactive events and alternatively of goodFor/badFor events. Later work by Wiebe's group as well as work by Ruppenhofer and Brandes (2015) speaks more generally of effect events.

[8] By contrast, in the first iteration of the shared task, we had counted a match when there was partial span overlap. In addition, we had used the Dice coefficient to assess the overlap between a system annotation and a gold standard annotation. Equally, for inter-annotator-agreement we had counted a match when there was partial span overlap.

Identification of Subjective Expressions

	LK_Run1	UDS_Run1	UDS_Run2	UDS_Run3	UDS_Run4	UDS_Run5	UDS_Run6*
system type		rule-based	rule-based	rule-based	rule-based	supervised	rule-based
f_1	0.350	0.239	0.293	0.346	0.346	**0.507**	0.351
p	0.482	0.570	0.555	0.564	0.564	**0.654**	0.572
r	0.275	0.151	0.199	0.249	0.249	**0.414**	0.253

Identification of Sources

	LK_Run1	UDS_Run1	UDS_Run2	UDS_Run3	UDS_Run4	UDS_Run5	UDS_Run6*
system type		rule-based	rule-based	rule-based	rule-based	supervised	rule-based
f_1	0.183	0.155	0.208	0.258	0.259	**0.318**	0.262
p	0.272	0.449	0.418	0.420	0.421	**0.502**	0.425
r	0.138	0.094	0.138	0.186	0.187	**0.233**	0.190

Identification of Targets

system type	LK_Run1	UDS_Run1	UDS_Run2	UDS_Run3	UDS_Run4	UDS_Run5	UDS_Run6*
		rule-based	rule-based	rule-based	rule-based	supervised	rule-based
f_1	0.143	0.184	0.199	0.253	0.256	0.225	**0.261**
p	0.204	**0.476**	0.453	0.448	0.450	0.323	0.440
r	0.110	0.114	0.127	0.176	0.179	0.173	**0.185**

Table 2: Full task: evaluation results based on the micro averages (results marked with a '*' are late submission)

2.6 Results

Two groups participated in our full task submitting one and six different runs respectively. Table 2 shows the results for each of the submitted runs based on the micro average of exact matches. The system that produced *UDS_Run1* presents a baseline. It is the rule-based system that UDS had used in the previous iteration of this shared task. Since this baseline system is publicly available, the scores for *UDS_Run1* can easily be replicated.

The rule-based runs submitted by UDS this year[9] (i.e *UDS_Run2*, *UDS_Run3*, *UDS_Run4* and *UDS_Run6*) implement several extensions that provide functionalities missing from the first incarnation of the UDS system:

- detection of grammatically-induced sentiment and the extraction of its corresponding sources and targets
- handling multiword expressions as subjective expressions and the extraction of their corresponding sources and targets
- normalization of dependency parses with regard to coordination

Please consult UDS's participation paper for details about these extensions of the baseline system.

The runs provided by Potsdam are also rule-based but they are focused on achieving generalization beyond the instances of subjective expressions and their sources and targets seen in the training data. They do so based on representing the relations between subjective expressions and their sources and targets in terms of paths through constituency parse trees. Potsdam's Run 1 (*LK_Run1*) seeks generalization for the subjective expressions already observed in the training data by merging all the paths of any two subjective expressions that share any path in the training data.

All the submitted runs show improvements over the baseline system for the three challenges in the full task with the exception of *LK_Run1* on target identification. While the supervised system used for Run 5 of the UDS group achieved the best results for the detection of subjective expressions and sources, the rule based system of UDS's Run 6 handled the identification of targets better.

When considering partial matches as well, the results on detecting sources improve only slightly, but show big improvements on targets with up to 25% points. A graphical comparison between exact and partial matches, can be found in Figure 1.

The results also show, that the poor f_1-measures can be mainly attributed to lacking recall. In other words, the systems miss a large portion of the manual annotations.

The two participating groups also submitted one and two runs for each of the subtasks respectively. Since the baseline system only supports the extraction of sources and targets according to the definition of the full task, a baseline score for the two subtasks could not be provided.

The results in Table 3 show improvements in

[9] The UDS systems have been developed under the supervision of Michael Wiegand, one of the workshop organizers.

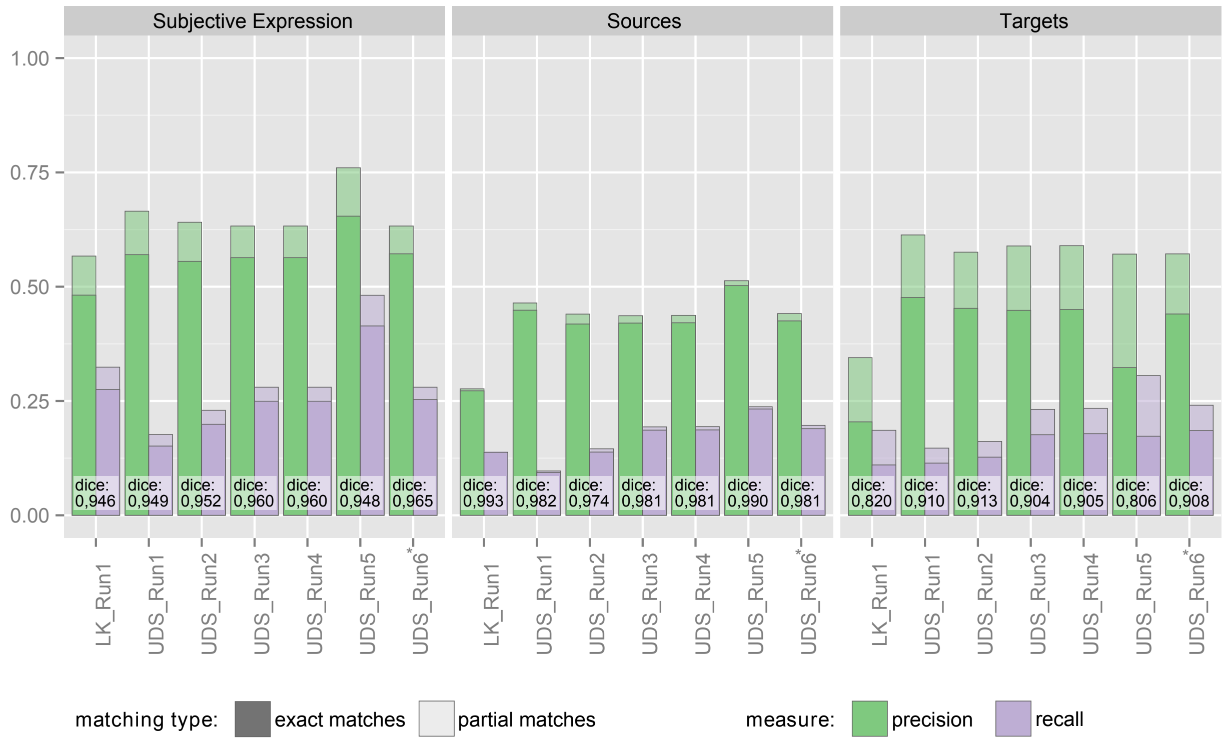


Figure 1: Comparison of exact and partial matches for the full task based on the micro average results (results marked with a '*' are late submissions)

both subtasks of about 15% points for the f_1-measure, when comparing the the best results between the full and the corresponding subtask. As in the full task, the identification of sources was best solved by a supervised machine learning system, when subjective expressions were given. The opposite is true for the target detection: The rule-based system outperforms the supervised machine learning system in the subtasks as it does in the full task.

The oberservations with respect to the partial matches are also constant across the full and the corresponding subtasks as can be seen in Figures 1 and 2: Target detection benefits a lot more than source detection when partial matches are considered as well.

3 Related Work

3.1 Other Shared Tasks

Many shared tasks have addressed the recognition of subjective units of language and, possibly, the classification of their polarity (SemEval 2013 Task 2, Twitter Sentiment Analysis (Nakov et al., 2013); SemEval-2010 task 18: Disambiguating sentiment ambiguous adjectives (Wu and Jin, 2010); SemEval-2007 Task 14: Affective Text (Strappar-

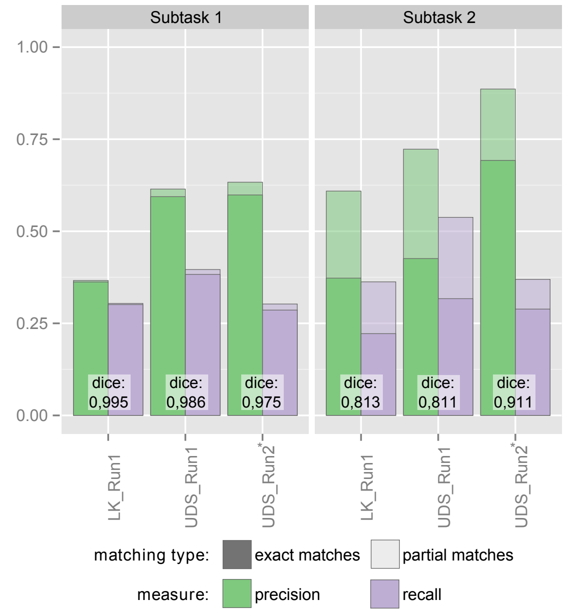


Figure 2: Comparison of exact and partial matches for the subtasks based on the micro average results (results marked with a '*' are late submission)

Subtask 1: Identification of Sources

	LK_Run1	UDS_Run1	UDS_Run2*
system type		supervised	rule-based
*f*1	0.329	**0.466**	0.387
p	0.362	0.594	**0.599**
r	0.301	**0.383**	0.286

Subtask 2: Identification of Targets

	LK_Run1	UDS_Run1	UDS_Run2*
system type		supervised	rule-based
*f*1	0.278	0.363	**0.407**
p	0.373	0.426	**0.692**
r	0.222	**0.317**	0.289

Table 3: Subtasks: evaluation results based on the micro averages (results marked with a '*' are late submissions)

ava and Mihalcea, 2007) *inter alia*).

Only of late have shared tasks included the extraction of sources and targets. Some relatively early work that is relevant to the task presented here was done in the context of the Japanese NTCIR[10] Project. In the NTCIR-6 Opinion Analysis Pilot Task (Seki et al., 2007), which was offered for Chinese, Japanese and English, sources and targets had to be found relative to whole opinionated sentences rather than individual subjective expressions. However, the task allowed for multiple opinion sources to be recorded for a given sentence if there were multiple expressions of opinion. The opinion source for a sentence could occur anywhere in the document. In the evaluation, as necessary, co-reference information was used to (manually) check whether a system response was part of the correct chain of co-referring mentions. The sentences in the document were judged as either relevant or non-relevant to the topic (=target). Polarity was determined at the sentence level. For sentences with more than one opinion expressed, the polarity of the main opinion was carried over to the sentence as a whole. All sentences were annotated by three raters, allowing for strict and lenient (by majority vote) evaluation. The subsequent Multilingual Opinion Analysis tasks NTCIR-7 (Seki et al., 2008) and NTCIR-8 (Seki et al., 2010) were basically similar in their setup to NTCIR-6.

While our shared task focussed on German, the most important difference to the shared tasks organized by NTCIR is that it defined the source and target extraction task at the level of individual subjective expressions. There was no comparable shared task annotating at the expression level, rendering existing guidelines impractical and necessitating the development of completely new guidelines.

Another more recent shared task related to STEPS is the Sentiment Slot Filling track (SSF) that was part of the Shared Task for Knowledge Base Population of the Text Analysis Conference (TAC) organised by the National Institute of Standards and Technology (NIST) (Mitchell, 2013). The major distinguishing characteristic of that shared task, which is offered exclusively for English language data, lies in its retrieval-like setup. In our task, systems have to extract all possible triplets of subjective expression, opinion source and target from a given text. By contrast, in SSF the task is to retrieve sources that have some opinion towards a *given* target entity, or targets of some *given* opinion sources. In both cases, the polarity of the underlying opinion is also specified within SSF. The given targets or sources are considered a type of *query*. The opinion sources and targets are to be retrieved from a document collection.[11] Unlike STEPS, SSF uses heterogeneous text documents including both newswire and discussion forum data from the Web.

3.2 Systems for Source and Target Extraction

Throughout the rise of sentiment analysis, there have been various systems tackling either target extraction (e.g. Stoyanov and Cardie (2008)) or source extraction (e.g. Choi et al. (2005), Wilson et al. (2005)). Only recently has work on automatic systems for the extraction of complete fine-grained opinions picked up significantly. Deng and Wiebe (2015a), as part of their work on opinion inference, build on existing opinion analysis systems to construct a new system that extracts triplets of sources, polarities, and targets from the MPQA 3.0 corpus Deng and Wiebe (2015b).[12] Their system extracts directly encoded opinions, that is ones that are not inferred but directly conveyed by lexico-grammatical means, as the basis for subsequent inference of implicit opinions. To extract explicit opinions, Deng and Wiebe (2015a)'s system incorporates, among others, a prior system by Yang and Cardie (2013) . That earlier system is trained to extract triplets of source span, opinion span and tar-

[10] NII [National Institute of Informatics] Test Collection for IR Systems

[11] In 2014, the text from which entities are to be retrieved is restricted to one document per query.

[12] Note that the specific (spans of the) subjective expressions which give rise to the polarity and which interrelate source and target are not directly modeled in Deng and Wiebe (2015a)'s task set-up.

get span, but is adapted to the earlier 2.0 version of MPQA, which lacked the entity and event targets available in version 3.0 of the corpus.[13]

A difference between the above mentioned systems and the approach taken here, which is also embodied by the system of Wiegand et al. (2014), is that we tie source and target extraction explicitly to the analysis of predicate-argument structures (and ideally, semantic roles), whereas the former systems and the corpora they evaluate against, are much less strongly guided by these considerations.

4 Conclusion and Outlook

We reported on the second iteration of the STEPS shared task for German sentiment analysis. Our task focused on the discovery of subjective expressions and their related entities in political speeches.

Based on feedback and reflection following the first iteration, we made a baseline system available so as to lower the barrier for participation in second iteration of the shared task and to allow participants to focus their efforts on specific ideas and methods. We also changed the evaluation setup so that a single reference annotation was used rather than matching against a variety of different references. This simpler evaluation mode provided participants with a clear objective function that could be learnt and made sure that the upper bound for system performance would be 100% precision/recall/F_1-score, whereas it was lower for the first iteration given that existing differences between the annotators necessarily led to false positives and negatives.

Despite these changes, in the end the task had only 2 participants. We therefore again sought feedback from actual and potential participants at the end of the IGGSA workshop in order to be able to tailor the tasks better in a future iteration.

Acknowledgments

We would like to thank Simon Clematide for helping us get access to the Swiss data for the STEPS task. For her support in preparing and carrying out the annotations of this data, we would like to thank Stephanie Köser.

We are happy to acknowledge the finanical support that the GSCL (German Society for Computational Linguistics) granted us in 2014 for the annotation of the training data, which served as test data in the first iteration of the IGGSA shared task.

Josef Ruppenhofer was partially supported by the German Research Foundation (DFG) under grant RU 1873/2-1.

Michael Wiegand was partially supported by the German Research Foundataion (DFG) under grant WI 4204/2-1.

References

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Pado. 2006. SALTO - A Versatile Multi-Level Annotation Tool. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 517–520.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Mark Davies and Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051.

Lingjia Deng and Janyce Wiebe. 2015a. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 179–189, Lisbon, Portugal, September. Association for Computational Linguistics.

Lingjia Deng and Janyce Wiebe. 2015b. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado, May–June. Association for Computational Linguistics.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125, Sofia, Bulgaria, August. Association for Computational Linguistics.

Wolfgang Lezius. 2002. TIGERsearch - Ein Suchwerkzeug für Baumbanken. In Stephan Busemann, editor, *Proceedings of KONVENS 2002*, Saarbrücken, Germany.

Margaret Mitchell. 2013. Overview of the TAC2013 Knowledge Base Population Evaluation: English

[13] Deng and Wiebe (2015a) also use two more systems that identify opinion spans and polarities but which either do not extract sources and targets at all (Yang and Cardie, 2014), or assume that the writer is always the source (Socher et al., 2013).

Sentiment Slot Filling. In *Proceedings of the Text Analysis Conference (TAC)*, Gaithersburg, MD, USA.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta and Georgia and USA. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Randolph Quirk, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Longman.

Josef Ruppenhofer and Jasper Brandes. 2015. Extending effect annotation with lexical decomposition. In *Proceedings of the 6th Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 67–76.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

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 NTCIR-6 Workshop Meeting*, pages 265–278.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2008. Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pages 185–203.

Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2010. Overview of Multilingual Opinion Analysis Task at NTCIR-8: A Step Toward Cross Lingual Opinion Analysis. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 209–220.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Veselin Stoyanov and Claire Cardie. 2008. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 817–824, Manchester, UK, August. Coling 2008 Organizing Committee.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74. Association for Computational Linguistics.

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 Gertrud Faaßand Josef Ruppenhofer, editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 174–184, Hildesheim, Germany, October. Universität Heidelberg.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yunfang Wu and Peng Jin. 2010. SemEval-2010 Task 18: Disambiguating Sentiment Ambiguous Adjectives. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 81–85, Stroudsburg and PA and USA. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2014. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 325–335, Baltimore, Maryland, June. Association for Computational Linguistics.

System documentation for the IGGSA Shared Task 2016

Leonard Kriese
Universität Potsdam
Department Linguistik
Karl-Liebknecht-Straße 24-25
14476 Potsdam
leonard.kriese@hotmail.de

Abstract

This is a brief documentation of a system, which was created to compete in the shared task on *Source, Subjective Expression and Target Extraction from Political Speeches (STEPS).* The system's model is created from supervised learning on using the provided training data and is learning a lexicon of *subjective expressions*. Then a slightly different model will be presented that generalizes a little bit from the training data.

1 Introduction

This is a documentation of a system, which has been created within the context of the IGGSA Shared Task 2016 STEPS[1]. Briefly, the main goal was to find *subjective expressions (SE)* that are functioning as opinion triggers, their *sources*, the originator of an SE, and their *targets*, the scope of an opinion. The system was aimed to perform on the domain of parliament speeches from the Swiss Parliament. The system's model was trained on the training data provided alongside the shared task and was from the same domain, preprocessed, with constituency parses from the Berkley Parser (Petrov and Klein, 2007) and had annotations of SEs and their respective targets and sources.

The model is using a mapping from grouped SEs to a set of "path-bundles", syntactic relations between SE, source and target. Since the learned SEs are a lexicon derived from the training data and are very domain-dependent, there will be a second model presented, which generalizes slightly from the training data by using the SentiWS (Remus et al., 2010) as a lexicon of SEs. There, the part-of-speech tag of each word from the SentiWS is mapped to a set of path-bundles.

2 System description

Participants were free in the way they could develop the system. They just had to identify subjective expressions and their corresponding target and source. Our system is using a lexical approach to find the subjective expressions and a syntactic approach in finding the corresponding target and source. First, all the words in the training data were lemmatized with the TreeTagger (Schmid, 1995), to keep the number of considered words as low as possible. Then the SE lexicon was derived from all the SEs in the training data. For each SE in the training data its path-bundle, a syntactic relation to the SE's target and source was stored. These path-bundles were derived from the constituency-parses from the Berkley Parser. For each sentence in the test data all the words were checked if they were SE candidates. If they were, their syntactic surroundings were checked as well. If these were also valid, a target and source was annotated. The test data was also lemmatized.

The outline of this paper is: the approach of deriving the syntactic relation of the SEs by introducing the concepts of "minimal trees" and "path-bundles" (Section 2.1 and Section 2.2) will be presented. Then the clustering of SEs and their path-bundles will be explained (Section 2.3) and a more generalized model (Section 2.4).

2.1 Minimal Trees

We use the term "minimal tree" for a sub-tree of a syntax tree given in the training data for each sentence with the following property: its root node is the least common ancestor of the SE, target and source[2]. From all the identified minimal trees so called path-bundles were derived. In Figure 1 and 2 you can see such minimal trees. These just focus on the part of the syntactic tree, which relates to

[1] Source, Subjective Expression and Target Extraction from Political Speeches(STEPS)

[2] Like the *lowest common multiple*, just for SE, target and source.

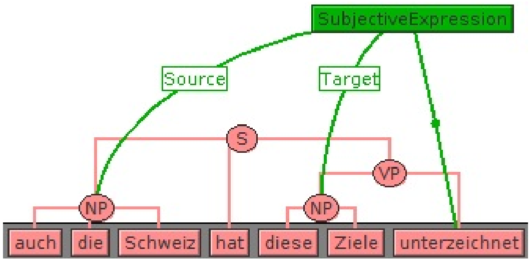


Figure 1: minimal tree covering *auch die Schweiz hat diese Ziele unterzeichnet*

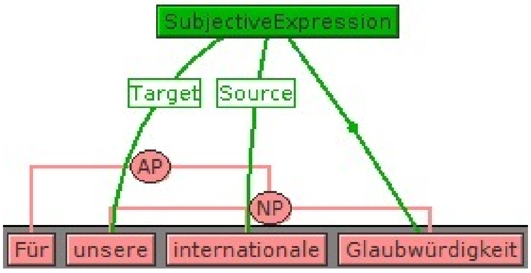


Figure 2: minimal tree covering *Für unsere internationale Glaubwürdigkeit*

SE, source and target. Following the root node of the minimal tree to the SE, target and source, path-bundles are extracted, as you can see in (1) and (2).

2.2 Path-Bundles

A path-bundle holds the paths in the minimal tree for the SE, target and source to the root node of the minimal tree.

(1) path-bundle for minimal tree in Figure 1
SE: [S, VP, VVPP]
T: [S, VP, NP]
S: [S, NP]

(2) path-bundle for minimal tree in Figure 2
SE: [NP, NN]
T: [NP, PPOSAT]
S: [NP, ADJA]

As you can see in (1) and (2), there is no distinction between terminals and non-terminals here, since SEs are always terminals (or a group of terminals) and targets and sources are sometimes one and the other. A path-bundle is expressing a syntactic relation between the SE, target and source and can be seen as a syntactic pattern. In practice many SEs have more than one path-bundle.
When the system is annotating a sentence, e.g. from the test data, and an SE is detected, the system checks the current syntax tree for one of the path-bundles an SE might have. If one bundle applies, source and target along with the SE will be annotated.
Also the flags, which appear in the training data, are stored to a path-bundle and will be annotated, when the corresponding path-bundle applies.

2.3 Clustering

After the training procedure every SE has its own set of path-bundles. To make the system more open to unseen data, the SEs were clustered in the following way: if an SE shared at least one path-bundle with another SE, they were put together into a cluster. The idea is, if SEs share a syntactic pattern towards their target and source, they are also syntactically similar and hence, should share their path-bundles. Rather than using a single SE mapped to a set of path-bundles, the system uses a mapping of a set of SEs to the set of their path-bundles.

(3) {befürworten, wünschen, nachleben, beschreiben, schützen, versprechen, verschärfen, erreichen, empfehlen, ausschreiben, verlangen, folgen, mitunterschreiben, beizustellen, eingreifen, appellieren, behandeln}

(4) {..., Regelung, Bodenschutz, Nichteintreten, Anreiz, Verteidigung, Kommentator, Kommissionsmotion, Verkehrsbelastung, Jugendstrafgesetz, Rückweisungsantrag, Konvention, Neutralitätsthematik, Europapolitik, Debatte,...}

(5) {lehnen ab, nehmen auf, ordnen an}

The clusters in (3), (4) and (5) are examples of what has been clustered in the training. This was done automatically and is presented here for illustration. As future work, we will consider manually merging some of the clusters and testing, whether that improves the performance.

2.4 Second model

The second model is generalizing a little bit from the lexicon of the training data, since the first model is very domain-dependent and should perform much worse on another domain than on the test data. The generalization is done by exchanging the lexicon learned from the training data with the words from the SentiWS (Remus et al., 2010).
This model is thus more generalized and not domain-dependent, but neither domain-specific. If

a word from the lexicon will be detected in a sentence, then all path-bundles, which begin with the same pos-tag, in the SE-path, will be considered for finding the target and source.

In general the sorting of the path-bundles is dependent from the leaf node in the SE-path, since the procedure is the following: find an SE and check if one of the path-bundles can be applied. Maybe, this can be done in a reverse way, where every node in a syntax tree is seen as a potential top-node of a path-bundle and if a path-bundle can be applied, SE, target and source will be annotated accordingly. This could be a heuristic for finding SEs without the use of a lexicon.

3 Results

In this part, the results of the two models, which ran on the STEPS 2016 data, will be presented.

Measure	Supervised	SentiWS
F1 SE exact	35.02	30.42
F1 source exact	18.29	15.62
F1 target exact	14.32	14.52
Prec SE exact	48.15	58.40
Prec source exact	27.23	34.66
Prec target exact	20.44	32.11
Rec SE exact	27.51	20.56
Rec source exact	13.77	10.08
Rec target exact	11.02	9.38

Table 1: Results of the system's runs on the main task.

Measure	Subtask A
F1 source exact	32.87
Prec source exact	36.23
Rec source exact	30.08
	Subtask B
F1 target exact	27.83
Prec target exact	37.29
Rec target exact	22.20

Table 2: Results of the system run on the subtasks.

The first system (Supervised) is the domain-dependent, supervised system with the lexicon from the training data and was the system, which was submitted to the IGGSA Shared Task 2016. The second system (SentiWS) is the system with the lexicon from the SentiWS. Speaking about Table 1, with the results for the main task, considering the F1-measure, the first system was better in finding SEs and sources but a little bit worse in finding targets.

The second system, the more general system, was better in the precision scores overall. This means, in comparison to the supervised system, that the classified SEs, targets and source were more correct. But it did no find as many as it should have found as the first system according to the recall scores. This leads to the assumption that the first system might overgenerate and is therefore hitting more of the true positives, but is also making more mistakes.

Looking at Table 2, the systemic approach is just different in terms of the lexicon of SEs and not in terms of the path-bundles. So there is no distinction between the two systems here, since all the SEs were given in the subtasks and only the learned path-bundles determined the outcome of the subtasks. For the system it seems easier to find the right sources, rather than the right targets, which is also proven by the numbers in Table 1.

4 Conclusion

In this documentation for the IGGSA Shared Task 2016 an approach was presented, which uses the provided training data. First, a lexicon of SEs was derived from the training data along with their path-bundles, indicating where to find their respective target and source. Two generalization steps were made by first, clustering SEs, which had syntactic similarities and second by exchanging the lexicon derived from the training data with a domain-independent lexicon, the SentiWS.

The first, very domain-dependent, system performed better than the more general second system according to the f-score. But the second system did not make as many mistakes in detecting SEs from the test data by looking at the precision score, so it might be worth to investigate into the direction of using a more general approach further.

The approach of deriving path-bundles from syntax trees itself is domain-independent, since it can be easily applied to any kind of parse. It would be nice to see, how the results will change, when other parsers, like a dependency parser, will be used. This is something for the future work.

References

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Tech-*

nologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

R. Remus, U. Quasthoff, and G. Heyer. 2010. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pages 1168–1171.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer.

Saarland University's Participation in the Second Shared Task on Source, Subjective Expression and Target Extraction from Political Speeches (STEPS-2016)

Michael Wiegand, Nadisha-Marie Aliman, Tatjana Anikina, Patrick Carroll, Margarita Chikobava, Erik Hahn, Marina Haid, Katja König, Leonie Lapp, Artuur Leeuwenberg, Martin Wolf, Maximilian Wolf
Spoken Language Systems, Saarland University
D-66123, Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de

Abstract

We report on the two systems we built for the second run of the shared task on *Source, Subjective Expression and Target Extraction from Political Speeches (STEPS)*. The first system is a rule-based system relying on a predicate lexicon specifying extraction rules for verbs, nouns and adjectives, while the second is a supervised system trained on the adjudicated test data of the previous run of this shared task.

1 Introduction

In this paper, we describe our two systems for the second run of the shared task on *Source, Subjective Expression and Target Extraction from Political Speeches (STEPS)* organized by the Interest Group on German Sentiment Analysis (IGGSA). In that task, both *opinion sources*, i.e. the entities that utter an opinion, and *opinion targets*, i.e. the entities towards which an opinion is directed, are extracted from German sentences. The opinions themselves have also to be detected automatically. The sentences originate from debates of the Swiss Parliament (*Schweizer Bundesversammlung*).

The first system is a rule-based system relying on a predicate lexicon specifying extraction rules for verbs, nouns and adjectives, while the second is a supervised classifier trained on the adjudicated test data of the previous edition of this shared task (Ruppenhofer et al., 2014).

2 Rule-based System

Our rule-based system is an extension of the rule-based system built for the first edition of this shared task as described in Wiegand et al. (2014). The pipeline of the rule-based system is displayed in Figure 1. The major assumption that underlies this system is that the concrete opinion sources and targets are largely determined by the opinion predicate[1] by which they are evoked. Therefore, the task of extracting opinion sources and targets is a lexical problem, and a lexicon for opinion predicates specifying the argument position of sources and targets is required. For instance, in Sentence (1), the sentiment is evoked by the predicate *liebt*, the source is realized by its subject *Peter* while the target is realized by its accusative object *Maria*.

(1) $[\text{Peter}]^{\textit{source}}_{\textit{subj}}$ **liebt** $[\text{Maria}]^{\textit{target}}_{\textit{obja}}$.
(Peter loves Maria.)

With this assumption, we can specify the demands of an opinion source/target extraction system. It should be a tool that given a lexicon with argument information about sources and targets for each opinion predicate

- checks each sentence for the presence of such opinion predicates,
- syntactically analyzes each sentence and
- determines whether constituents fulfilling the respective argument information about sources and targets are present in the sentence.

In the following, we briefly describe the linguistic processing (Section 2.1) and the mechanism for extracting rules (Section 2.2). Then, we introduce the extensions we applied for this year's submission (Section 2.3). For general information regarding the architecture of the system, we refer the reader to Wiegand et al. (2014).

The version of the rule-based system that has been revised for this year's shared task has been made publicly available[2] allowing researchers to

[1] We currently consider opinion verbs, nouns and adjectives as potential opinion predicates.

[2] https://github.com/miwieg/german-opinion-role-extractor

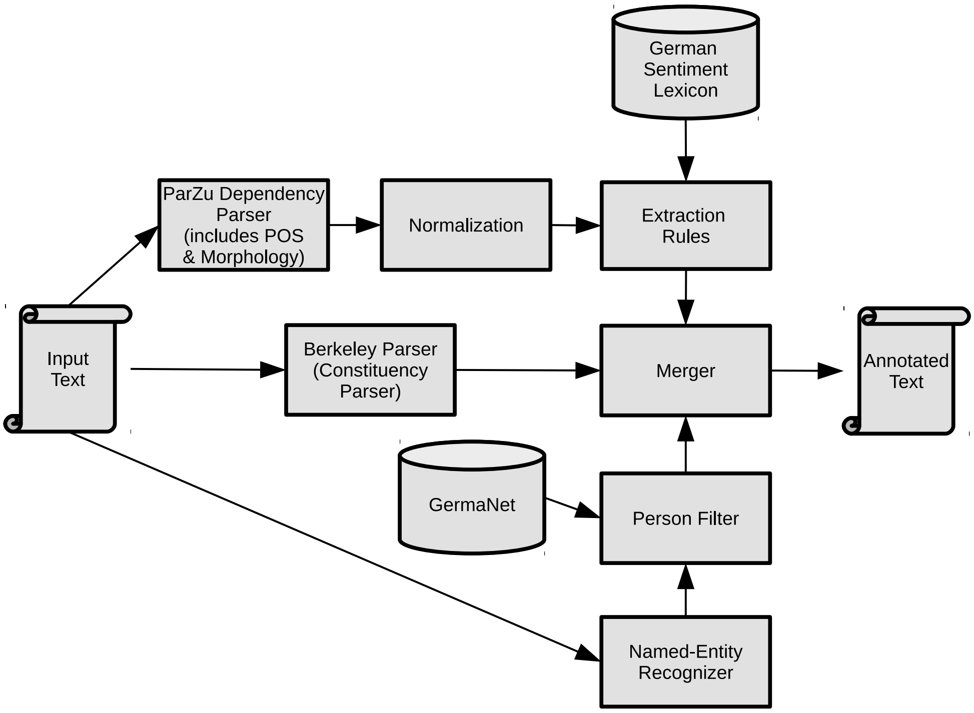


Figure 1: Processing pipeline of the rule-based system.

test different sentiment lexicons with different argument information about opinion sources and targets.

2.1 Linguistic Processing

Even though the data for this task already come in a parsed format, we felt the need to add further linguistic information. In addition to the existing constituency parse provided by the Berkeley parser (Petrov et al., 2006), we also included dependency parse information. With that representation, relationships between opinion predicates and their sources and targets can be formulated more intuitively.[3]

As a dependency parser, we chose *ParZu* (Sennrich et al., 2009). We also carried out some normalization on the parse output in order to have a more compact representation. To a large extent, the type of normalization we carry out is in line with the output of dependency parsers for English, such as the Stanford parser (de Marneffe et al., 2006). It is included since it largely facilitates writing extraction rules. The normalization includes

(a) active-passive normalization

(b) conflating several multi-edge relationships to one-edge relationships

(c) particle-verb reconstruction

These normalization steps are explained in more detail in Wiegand et al. (2014).

We also employed a semantic filter for the detection of opinion sources. Since such entities can only represent persons or groups of persons, we employed a named-entity recognizer (Benikova et al., 2015) to recognize person names and GermaNet (Hamp and Feldweg, 1997) to establish that a common noun represents a person or a group of persons.

2.2 The Extraction Rules

The heart of the rule-based system is a lexicon that specifies the (possible) argument positions of sources and targets. So far, there does not exist any lexicon with that specific information which is why we came up with a set of default rules for the different parts of speech. The set of opinion predicates are the subjective expressions from the PolArt system (Klenner et al., 2009). Every mention of such expressions will be considered as a mention of an opinion predicate, that is, we do not carry out any subjectivity word-sense disambiguation (Akkaya et al., 2009).

[3] As a matter of fact, the most appropriate representation for that task is semantic-role labeling (Ruppenhofer et al., 2008; Kim and Hovy, 2006; Wiegand and Klakow, 2012), however, there currently do not exist any robust tools of that kind for German.

These default extraction rules are designed in such a way that for a large fraction of opinion predicates with the pertaining part of speech they are correct. The rules are illustrated in Table 1. We currently have distinct rules for verbs, nouns and adjectives. All rules have in common that for every opinion predicate mention, at most one source and at most one target is assigned. The rules mostly adhere to the dependency relation labels of ParZu.[4]

The rule for verbs assumes sources in subject and targets in object position (1). Note that for targets, we specify a priority list. That is, the most preferred argument position is a dative object (*objd*), the second most preferred position is an accusative object (*obja*), etc. In computational terms, this means that the classifier checks the entire priority list (from left to right) until a relation has matched in the sentence to be classified. For prepositional complements, we also allow a wildcard symbol (*pobj-**) that matches all prepositional complements irrespective of its particular head, e.g. *über das Freihandelsabkommen* (*pobj-ueber*) in (2).

(2) [Deutschland und die USA]$^{source}_{subj}$ **streiten** [über das Freihandelsabkommen]$^{target}_{pobj-ueber}$.
(Germany and the USA quarrel over the free trade agreement.)

For nouns, we allow determiners (possessives) (3) and genitive modifiers (4) as opinion sources whereas targets are considered to occur as prepositional objects.

(3) [Sein]$^{source}_{det}$ **Hass** [auf die Regierung]$^{target}_{pobj-auf}$ ...
(His hatred towards the government ...)

(4) Die **Haltung** [der Kanzlerin]$^{source}_{gmod}$ [zur Energiewende]$^{target}_{pobj-zu}$ ...
(The chancellor's attitude towards the energy revolution ...)

The rule for adjectives is different from the others since it assumes the source of the adjective to be the speaker of the utterance. Only the target has a surface realization. Either it is an attributive adjective (5) or it is the subject of a predicative adjective (6).

Part of Speech	Source	Target
verb	subj	objd, obja, objc, obji, s, objp-*
noun	det, gmod	objp-*
adjective	*author*	attr-rev, subj

Table 1: Extraction rules for verb, noun and adjective opinion predicates.

(5) Das ist ein [guter]$^{target}_{attr-rev}$ **Vorschlag**.
(This is a good proposal.)

(6) [Der Vorschlag]$^{target}_{subj}$ ist **gut**.
(The proposal is good.)

Our rule-based system is designed in such a way that, in principle, it would also allow more than one opinion frame to be evoked by the same opinion predicate. For example, in *Peter überzeugt Maria/Peter convinces Maria*, one frame sees *Peter* as source and *Maria* as target, and another frame where the roles are switched. Our default rules do not include such cases, since such property is specific to particular opinion predicates.

2.3 Extensions

In this subsection, we present the extensions we added to the existing rule-based system from the previous iteration of this shared task.

2.3.1 Partial Analysis

Our system has been modified in such a way that it can now accept a partial analysis as input and process it further. By that we mean the existing annotation of subjective expressions as specified by the subtask of this shared task. Given such input, the system just assigns sources and targets for these existing expressions. (We also implemented another mode in which the opinion predicates according the given sentiment lexicon would additionally be recognized including their opinion roles.) Opinion predicates are typically ambiguous; our lexicon-based approach is therefore limited. This is a well-known and well-researched problem. On the other hand, the task of extracting opinion sources and targets given some opinion predicates is a completely different task, which is comparatively less well researched. Our mode allowing partial analysis as input should allow researchers interested in opinion role extraction to have a suitable test bed without caring for the detection of subjectivity.

[4] The definition of those dependency labels is available at `https://github.com/rsennrich/ParZu/blob/master/LABELS.md`

2.3.2 Grammatically-Induced Sentiment

An important aspect of opinion-role extraction that was completely ignored in the initial version of the rule-based system is the sentiment that is not evoked by common opinion predicates but sentiment that is evoked by certain grammatical constructions. We focus on certain types of modalities (7) and tenses (8). Such type of sentiment is detected without our extraction lexicon (§2.2).

(7) [Kinder **sollten** nicht auf der Straße spielen.]^{target} *source: speaker*
(Children should not play on the street.)

(8) [Er **wird** mal ein guter Lehrer sein.]^{target} *source: speaker*
(He is going to become a good teacher.)

(9) Der Puls des Patienten *wird* gemessen.
(The patient's pulse is measured.)

It is triggered by certain types of lexical units, that is, modal verbs, such as *sollte*, or auxiliary verbs, such as *werden*. However, unlike the lexical units from our extraction lexicon, some of these verbs require some further disambiguation. For instance, the German auxiliary *werden* is not exclusively used to indicate future tense as in (8) but it is also used for building passive voice (9). Therefore, our module carries out some contextual disambiguation of these words.

Grammatically-induced sentiment also systematically differs from lexical sentiment in the way in which opinion roles are expressed. While for lexical sentiment, the argument position of the sources and targets is dependent on the specific lexical unit that conveys the sentiment and therefore has to be specified by lexical rules, the types of grammatically-induced sentiment that we cover share the same argument positions for sources and targets. Typically, the source is the speaker of the utterance and the target is the entire sentence in which the tense or modal occurs. Of course, in case of compound sentences, the scope of the target is only restricted to the clause in which the auxiliary/modal verb occurs (10).

(10) [Er **wird** mal ein guter Lehrer sein]^{target}, da er gut erklären kann. *source: speaker*
(He is going to become a good teacher since he can explain things well.)

2.3.3 Morphological Analysis

Opinion sources are typically persons or groups of persons. In order to ensure that only NPs that match this semantic type are classified as sources, we employed a semantic filter that used the prediction of a named-entity recognizer in case of proper nouns and GermaNet (Hamp and Feldweg, 1997) in case of common nouns. The latter approach, however, is limited considering the high frequency of compounding in German. We observed that in case an opinion source was represented by a compound, such as *SPD-Landtagsabgeordneter*, it could not be established as a person since that term was not included in GermaNet. We examined whether this coverage problem could be solved by morphologically analyzing those compounds and then only looking up their heads (e.g. *Abgeordneter*) which are more likely to be included in GermaNet. A similar experiment was carried out to match opinion predicates in compounds (e.g. *Frühjahrs<u>aufschwung</u>* or *Telefon<u>terror</u>*). Our initial experiments with *morphisto* (Zielinski and Simon, 2009), however, showed no improvement in either opinion source extraction or subjectivity detection due to the high ambiguity in noun compound structures.

2.3.4 Normalizing Conjunctions

In the original version of our system we already incorporated a set of normalization steps of simplifying the dependency parse (Wiegand et al., 2014). The result was a more compact representation of sentences that abstracts from the surface realization of a sentence. This made it simpler to state extraction rules for the extraction of opinion sources and targets. In our submission for this year's task, we added a further normalization step dealing with conjunctions. The original dependency parse typically only directly connects one conjunct with the syntactic roles relevant for opinion roles. For instance, in Figure 2(a) only *lügt* is connected with *Er* by a subject relation. Therefore, our original system would only be able to establish that *Er* is some opinion role of *lügt*. In such cases, we also add another edge with the subject relation connecting the second conjunct (*betrügt*) and its subject (Figure 2(b)).

We also incorporate a set of rules to handle coordination for predicative and attributive adjectives.[5]

[5] For nouns, we could not figure out unambiguous relations where adding further edges would have increased the extraction of sources or targets.

While for predicative adjectives, the subjective relation has to be duplicated, for attributive adjectives, the edge *attr* needs to be duplicated (see Figure 3).

2.3.5 Alternative NLP Tools

We critically assessed the choice of NLP tools used in our original system and compared them with alternatives.

As far as both constituency and dependency parsing is concerned, it was not possible for us to find more effective alternatives. Either the corresponding parser could not be converted in our existing format so that the system would still work as before, or, the parsing output was notably inferior and produced worse extraction performance when incorporated in our processing pipeline.

As far as named-entity recognition is concerned, we replaced the original tagger, the German model from the Stanford named-entity tagger (Faruqui and Padó, 2010), by a more recent tagger, i.e. GermaNER (Benikova et al., 2015). We primarily decided in favour of this replacement since the Stanford named-entity tagger occasionally overrides the given sentence-boundary detection.

2.4 Multiword Expressions

Not every opinion predicate is a unigram token. So far, the only *multiword expressions* conveying sentiment that our system was able to process were phrasal verbs, as *gab auf* in (11).

(11) Er **gab** das Rauchen vor 10 Jahren **auf**.
(He gave up smoking 10 years ago.)

We modified our system in such a way that extraction rules can now also be specified for arbitrary multiword expressions. Matching multiword expressions in German sentences is not trivial since

- multiword expressions can be discontinuous sequences of tokens (e.g. (12), (13)),
- the order of tokens between the canonical form and mentions in specific sentences may vary (e.g. (13)), and
- several tokens between the canonical form and mentions in specific sentences may differ (e.g. reflexive pronoun *sich* in (12)).

In order to account for these properties, our matching algorithm considers the dependency parse of a sentence. We identify a multiword expression if the tokens of a particular expression are all directly connected via edges in a dependency parse. The multiword expressions must hence form a connected subgraph of the parse in which all tokens of the multiword expression and only those are included.

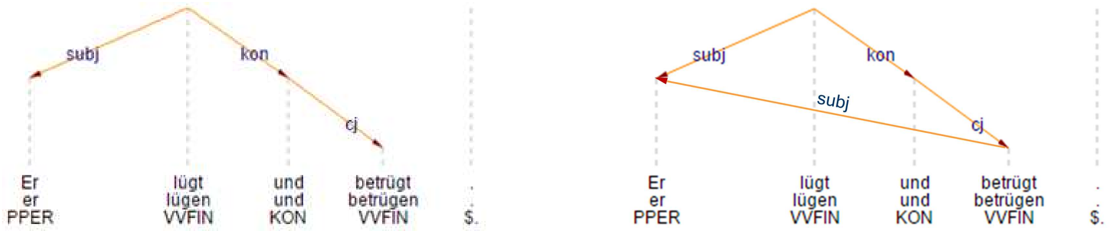
(12) *sich benehmen wie die Axt im Walde* (*act like a brute*):
Er sagte, dass ich **mich benehmen** würde, **wie die Axt im Walde**.
(He told me that I acted like a brute.)

(13) *sein wahres Gesicht zeigen* (*to show one's true colours*):
Unter Alkoholeinfluss **zeigte** er **sein wahres Gesicht**.
(Under the influence of alcohol, he showed his true colours.)

Since we are not aware of any publicly available lexicon for multiword expressions, we extract them automatically from a German corpus. For that, we use the parsed *deWaC* (Baroni et al., 2009). We focus on those multiword expressions that follow a systematic pattern. We chose *reflexive verbs* (e.g. *sich freuen*, *sich schämen*, *sich fürchten*) and *light-verb constructions* (e.g. *Angst haben*, *Kummer machen*, *Acht geben*). In order to extract reflexive verbs, we extracted opinion verbs frequently occurring with a reflexive pronoun (we restrict the pronoun to be the accusative object of the verb). In order to extract light-verb constructions, we first manually selected a set of common light verbs (e.g. *haben*, *machen*, *geben*) and then looked for opinion nouns that often co-occur with these light verbs (we restrict the opinion noun to be the accusative object of the light verb). In total, we thus extracted about 4700 multiword expressions.

3 Supervised System

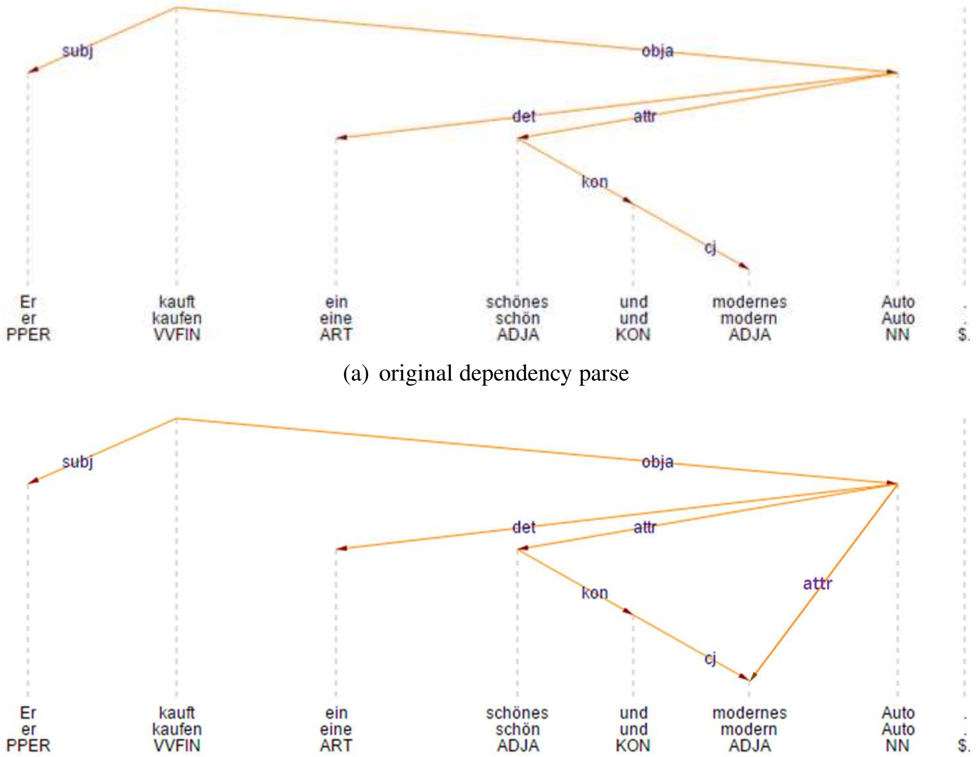
Since we participated in the second edition of this shared task, it meant that we were able to exploit the manually-labeled test data of the previous shared task (Ruppenhofer et al., 2014) as training data for a supervised classifier. In the previous edition, also a supervised system was submitted, however, it only considered as labeled training data texts automatically translated from English to German. Moreover, only opinion sources were con-



(a) original dependency parse

(b) normalized dependency parse

Figure 2: Illustration of normalizing dependency parses with verb coordination.



(a) original dependency parse

(b) normalized dependency parse

Figure 3: Illustration of normalizing dependency parses with adjective coordination.

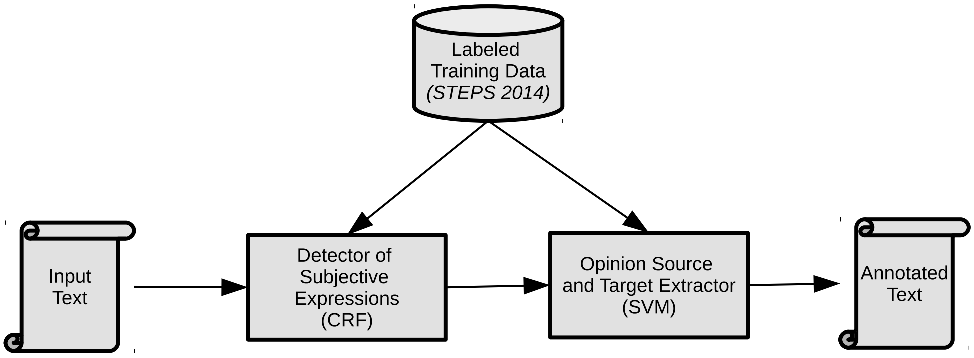


Figure 4: Processing pipeline of the supervised system.

Type	Feature Templates
words	unigram features: target word and its two predecessors/successors bigrams features: bigrams of neighboring words from unigram features
part of speech	unigram features: part-of-speech tag of target word and its two predecessors/successors bigram features: bigrams of neighboring part-of-speech tags from unigram features bigram features: trigrams of neighboring part-of-speech tags from unigram features
sentiment lexicon	is either of the words (window is that of the unigram features) an opinion predicate according to sentiment lexicon

Table 2: Feature templates employed for the CRF classifier to detect subjective expressions.

sidered. We believe that considering actual German text presents a much higher quality of training data than text that has automatically been translated into German.

The processing pipeline of our supervised system is illustrated in Figure 4. For this approach, we employed the same NLP tools as in our rule-based system in order to ensure comparability.

Our supervised system comprises two classifiers: The first is to detect opinion predicates. For that, we employ a conditional random field (Lafferty et al., 2001). As an implementation, we chose *CRF++*[6]. As a motivation, we chose a sequence-labeling algorithm because the task of detecting opinion predicates is similar to other tagging problems, such as part-of-speech tagging or named-entity recognition. The feature templates for our sentiment tagger are displayed in Table 2. We use CRF++ in its standard configuration; as a labeling scheme, we used the simple IO-notation.

The second classifier extracts for an opinion predicate detected by the CRF the corresponding opinion source or target, if they exist. For this support vector machines (SVM) were chosen. As an implementation, we used SVM^{light} (Joachims, 1999). The instance space is a set of tuples comprising candidate opinion roles and opinion predicates (detected by the previous sentiment detection). We use different sets of candidate phrases for opinion sources and opinion targets. For opinion sources, the set of candidates is the set of noun phrases in a sentence. Opinion sources are typically persons or groups of persons and, therefore, only noun phrases are eligible to represent such opinion role. Opinion targets, on the other hand, cannot be reduced to one semantic type. Targets can be various types of entities, both animate and inanimate. They can even represent entire propositions. As a consequence, we consider every constituent phrase of a sentence as a candidate opinion target.

SVM were chosen as a learning method since this task deals with a more complex instance space, and SVM, unlike sequence labelers, allow a fairly straightforward encoding of that instance space. The features we employed for this classifier are illustrated in Table 3.

4 Experiments

In this section, we evaluate the 7 runs officially submitted to the shared task. Table 4 displays the different properties of the different runs. The first 5 runs are rule-based systems, while the last run is a supervised system. *Rule-Based-2014* is the best rule-based system run in the previous iteration of this shared task (Wiegand et al., 2014) using the PolArt-sentiment lexicon (Klenner et al., 2009). *Rule-Based-2016-plain* is as *Rule-Based-2014* with various bugs removed. *Rule-Based-2016-gram* is as *Rule-Based-2016* with the module on grammatically-induced sentiment analysis (Section 2.3.2) switched on. *Rule-Based-2016-conj* is as *Rule-Based-2016-gram* but also with normalization of conjunctions (Section 2.3.4) switched on. The last system, *Supervised* is the supervised classifier presented in Section 3.

Table 5 displays the (micro-average) performance (exact matches) of the different configurations on the full task. **SE** evaluates the detection of subjective expressions, **Source** the detection of opinion sources and **Target** the detection of opinion targets.

Table 5 shows that the extensions made to the 2014-system result in some improvement. This improvement is caused by a notable rise in recall. The normalization of conjunctions and the treatment of multiword expressions only produce mild performance increases. We assume that this is due to the fact that, in the test data, there are only few cases of the conjunctions we deal with and also only few cases of the multiword expressions we extracted from a corpus. If one compares the rule-based systems with the supervised

[6] `https://code.google.com/p/crfpp/`

Type	Features
candidate opinion role	phrase label of candidate opinion role (e.g. *NP*, *VP*, *SBAR* etc.) lemma of head of phrase representing candidate opinion role part-of-speech of head of phrase representing candidate opinion role is head of phrase representing candidate opinion role some named entity? is candidate opinion role at the beginning of the sentence?
opinion predicate	lemma of opinion predicate part-of-speech of opinion predicate
relational	distance between opinion role candidate and opinion predicate dependency path from opinion role candidate to opinion predicate part-of-speech label of head of opinion role candidate and opinion predicate phrase label of opinion role candidate and part-of-speech tag of head of opinion predicate

Table 3: Features employed for the SVM classifier to extract opinion sources and targets.

Run	Properties
Rule-Based-2014	previous system (Wiegand et al., 2014) as it was publicly available
Rule-Based-2016-plain	as Rule-Based-2014 with various bugs removed
Rule-Based-2016-gram	Rule-Based-2016 with module on grammatically-induced sentiment analysis (Section 2.3.2) switched on
Rule-Based-2016-conj	as Rule-Based-2016-gram but also with normalization of conjunctions (Section 2.3.4) switched on
Rule-Based-2016-mwe	as Rule-Based-2016-conj but also with additional multiword expressions as part of the sentiment lexicon (Section 2.4)
Supervised	supervised learning system as discussed in Section 3

Table 4: The different properties of the different runs.

Run	Measure	SE	Source	Target
Rule-Based-2014	Prec	57.01	44.85	47.64
	Rec	15.14	9.70	11.41
	F	23.93	15.50	18.41
Rule-Based-2016-plain	Prec	55.54	41.83	45.27
	Rec	19.90	13.83	12.71
	F	29.30	20.79	19.85
Rule-Based-2016-gram	Prec	56.36	42.04	44.83
	Rec	24.95	18.62	17.63
	F	34.59	25.81	25.30
Rule-Based-2016-conj	Prec	56.36	42.11	45.01
	Rec	24.95	18.67	17.85
	F	34.59	25.88	25.57
Rule-Based-2016-mwe	Prec	57.20	45.52	44.03
	Rec	25.32	18.95	18.53
	F	35.10	26.22	**26.08**
Supervised	Prec	65.42	50.24	32.31
	Rec	41.41	23.25	17.29
	F	**50.72**	**31.79**	22.52

Table 5: Evaluation of the different runs of the Main Task.

Run	Task	Measure	SE	Source	Target
Rule-Based-2016-mwe	Full Task	Prec	57.20	45.52	44.03
		Rec	25.32	18.95	18.53
		F	35.10	26.22	**26.08**
	Subtask	Prec	100.0	59.86	69.24
		Rec	100.0	28.60	28.87
		F	100.0	38.70	**40.75**
Supervised	Full Task	Prec	65.42	50.24	32.31
		Rec	41.41	23.25	17.29
		F	50.72	31.79	22.52
	Subtask	Prec	100.0	59.40	42.60
		Rec	100.0	38.29	31.69
		F	100.0	**46.57**	36.35

Table 6: Evaluation of the Subtask.

system, one notices that on the detection of subjective expressions, the supervised system largely outperforms the rule-based system. Both precision and recall are improved. Obviously, the supervised system is the only classifier capable of disambiguating subjective expressions (Akkaya et al., 2009). Moreover, it seems to detect more subjective expressions than are contained in a common sentiment lexicon (which is the backbone of the subjectivity detection of the rule-based system). The supervised system, however, is less effective on the extraction on targets. Targets are more difficult to extract than sources in general since they can be more heterogeneous linguistic entities. Sources, for instance, are typically realized as noun phrases, whereas targets can be different types of phrases. We also observed that due to the fact that many targets are comparably large spans, parsing errors also affect this type of opinion roles more frequently. On the other hand, the constituents typically representing sources, i.e. (small) noun phrases, can be correctly recognized more easily. The supervised system may also outperform the rule-based system on the extraction of sources, since it can memorize certain entities with a high prior likelihood to be sources. For instance, first person pronouns (e.g. *I* or *we*) are very likely candidates for sources. This type of information cannot be incorporated in the rule-based classifier.

Table 6 compares the best rule-based system (i.e. *System-2016-mwe*) and the supervised system on both the full task and the subtask (again: micro-average performance – exact matches). In the subtask, subjective expressions are already given and only sources and targets have to be extracted. Obviously the subtask is easier which can be seen by the notably higher performance scores on both source and target extraction for both approaches. As on the full task, on the extraction of sources the supervised system outperforms the rule-based system, while on the extraction of targets, the rule-based system outperforms the supervised system.

5 Conclusion

We reported on the two systems we devised for the second edition of the shared task on *Source, Subjective Expression and Target Extraction from Political Speeches (STEPS)*. The first system is a rule-based system relying on a predicate lexicon specifying extraction rules for verbs, nouns and adjectives, while the second is a supervised classifier trained on the adjudicated test data of the previous edition of this shared task.

The supervised classifier scores well on the detection of subjective expressions and opinion sources. The rule-based system produces the best scores for the extraction of targets. Given the general low performance scores, we assume that the task of opinion source and target extraction still requires some further research.

Acknowledgements

This work was partially supported by the German Research Foundation (DFG) under grant WI 4204/2-1.

References

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–199, Singapore.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetti. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Darina Benikova, Seid Muhie Yimam, Prabhakaran Santhanam, and Chris Biemann. 2015. GermaNER: Free Open German Named Entity Recognition Tool. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 31–38, Essen, Germany.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 449–454, Genoa, Italy.

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

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.

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.

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.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. PolArt: A Robust Tool for Sentiment Analysis. In *Proceedings of the Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 235–238, Odense, Denmark.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the International Conference on Machine Learning (ICML)*, Williamstown, MA, USA.

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, Swapna Somasundaran, and Janyce Wiebe. 2008. Finding the Source and Targets of Subjective Expressions. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 2781–2788, Marrakech, Morocco.

Josef Ruppenhofer, Roman Klinger, Julia Maria Struß, Jonathan Sonntag, and Michael Wiegand. 2014. 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.

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, German.

Michael Wiegand and Dietrich Klakow. 2012. Generalization Methods for In-Domain and Cross-Domain Opinion Holder Extraction. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 325–335, Avignon, France.

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.

Andrea Zielinski and Christian Simon. 2009. Morphisto – An Open Source Morphological Analyzer for German. In *Proceedings of the 2009 Conference on Finite-State Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP 2008*, pages 224–231. IOS Press Amsterdam, The Netherlands.