Published in: Ruppenhofer, Josef/Struß, Julia Maria/Wiegand, Michael (eds.): IGGSA Shared Task on Source and Target Extraction from Political Speeches. Bochum: Ruhr-Universität Bochum, 2016. p. 1-9.
(Bochumer Linguistische Arbeitsberichte 18)

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.

[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.

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]

[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.

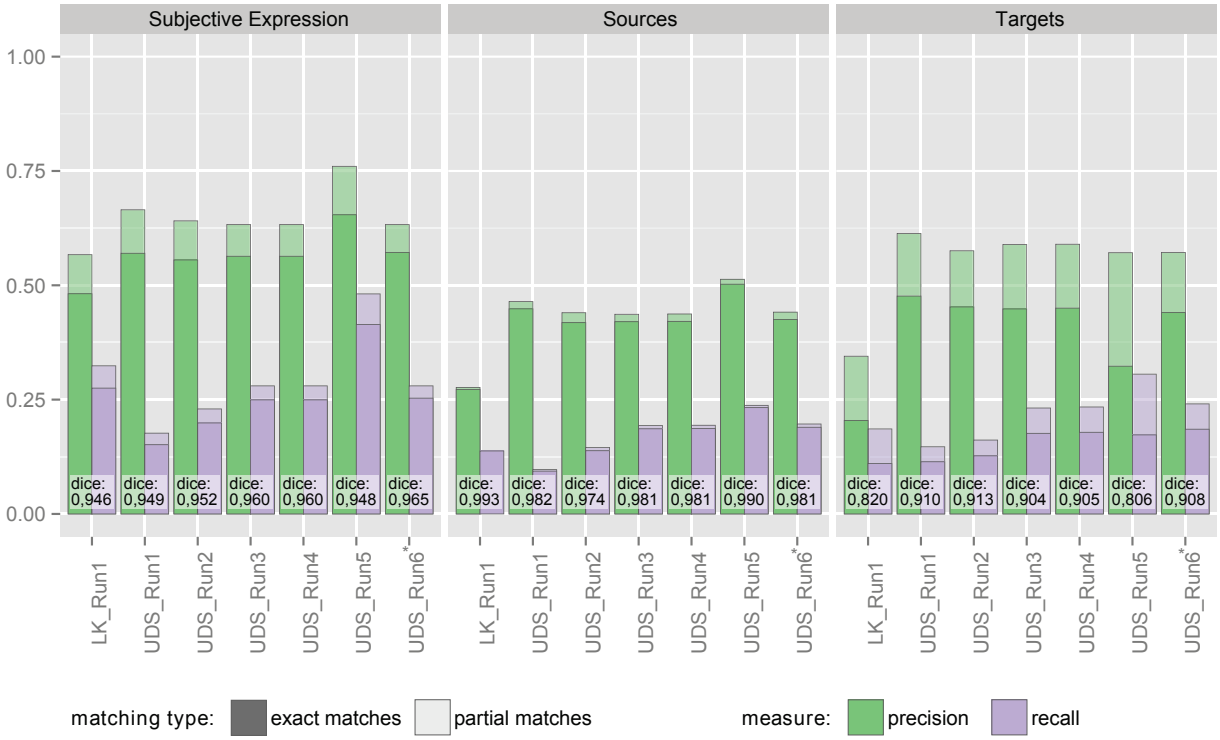

Figure 1: Comparison of exact and partial matches for the full task based on the micro average results (results marked with a '*' are late submissions)

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.

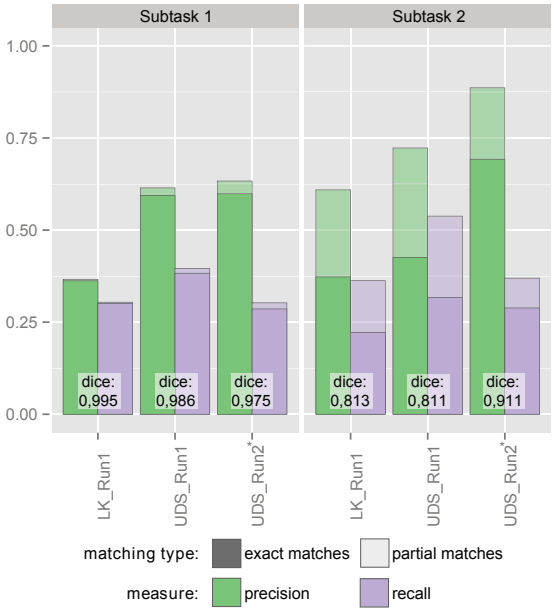

Figure 2: Comparison of exact and partial matches for the subtasks based on the micro average results (results marked with a '*' are late submission)

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-

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 lexicogrammatical 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.

[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).

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

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.