Josef Ruppenhofer, Julia Maria Struß, Jonathan Sonntag, Stefan Gindl

IGGSA-STEPS: Shared Task on Source and Target Extraction from Political Speeches

Accurate opinion mining requires the exact identification of the source and target of an opinion. To evaluate diverse tools, the research community relies on the existence of a gold standard corpus covering this need. Since such a corpus is currently not available for German, the Interest Group on German Sentiment Analysis decided to create such a resource and make it available to the research community in the context of a shared task. In this paper, we describe the selection of textual sources, development of annotation guidelines, and first evaluation results in the creation of a gold standard corpus for the German language.

1 Introduction

Opinion source and target extraction is the area of opinion mining aiming at identifying the source (i.e., whose opinion?) as well as the target (i.e., about what?) of an opinion. It is applicable to free language texts, where this kind of information cannot be derived from meta-data. Source and target extraction turns out to be a surprisingly difficult task. Intuitively, humans should be easily capable of accomplishing it, yet they often founder on the subtleties of language. While a brief glance at a text gives the impression of an easily solvable task, delving into it reveals its complexity. A varying number of sources/targets might confuse the reader, in other cases the source/target might not be present in the sentence, or it is difficult to decide on the linguistic span of the source/target. A task so difficult to solve for humans poses an even bigger challenge for computers. With their at most limited understanding of human language, solving such a task requires sophisticated algorithms. This is aggravated by the fact that the data publicly available for machine learning purposes is too sparse.

The paper we present here summarizes the efforts of the Interest Group of German Sentiment Analysis (IGGSA)[1] to create a publicly available resource serving as a gold standard corpus for opinion source and target extraction. The corpus consists of a large number of speech transcripts from debates in the Swiss parliament and contains annotations for the evaluation of source and target extraction systems. IGGSA plans to use the corpus as part of a shared task focusing on source and target extraction from political speeches (STEPS) in the run-up to the KONVENS conference 2014 in Hildesheim.

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

In this paper we discuss the choice of Swiss parliament speeches, report the details of the annotation guidelines and show evaluation results of a first round of manual annotations.

2 Related Work

An important aspect of opinion mining systems is their ability to establish a connection between subjective expressions and their sources and targets. A system capable of doing this provides a holistic picture of an expressed opinion. Sentiment analysis systems must be able to reliably tie opinions or subjective states to their sources and targets. This is a non-trivial task as some sentiment-bearing expressions are not linked to the sources, and some not even to the targets, of opinion. In the best case, the source and target correspond to semantic roles of sentiment-bearing predicates that can be expressed as syntactic arguments (Ruppenhofer et al., 2008). For instance, the subject of *love* in (1) is the source of the positive sentiment expressed and the object is the target of the sentiment.

(1) I really LOVE the players and the staff. [www]

However, a direct tie-in with semantic role labeling is usually not the chosen way of handling the extraction of sources and targets. In the following, we discuss the reasons for this and some of the alternative problem statements that have been adopted.

2.1 Attribution and nesting of sources

In the case of one important sub-class of sentiment-bearing expressions, called expressive subjective elements by Wiebe et al. (2005), a grammatical link exists between the opinion expression and the target[2], but not necessarily to the source. For instance, in the case of *idiotic* we know that what the adjective modifies or is predicated of is the target of the sentiment conveyed. Thus, "exit" is the target in (2) and "[t]hat" is the target in (3). Note, however, that the sources differ between the two examples: in (2), the source is the writer of the text, whereas in (3) it is the quoted speaker Irvine.

(2) His rude, crude response and IDIOTIC exit from his duties is hardly deserving of the praise he has attracted. [www]

(3) "That was IDIOTIC," Irvine told talkSPORT . [www]

Rather than connect expressions of opinion only to their immediate sources, it is desirable to keep track of the chain of transmission. In the MPQA-corpus (Wiebe et al., 2005), for instance, levels of nesting are recorded that would show for a sentence like (3) that not only is Irvine the source of the opinion expressed by *idiotic* but that we come to know this only via an utterance of the writer of the text in which Irvine's speech is presented. In the annotations we produce, nesting is not explicitly marked but can be reconstructed from the annotations, as discussed in Section 3.3.

[2]This link may either take the form of a predicate-argument or a modifier-head relationship.

2.2 Definitions of target

The main issue with respect to targets is whether the analysis should address only what one may call "local" targets, that is expressions that are semantic valents and syntactic dependents of a particular sentiment-bearing predicate, or whether it should also take into account other targets that are pragmatically relevant. To illustrate the difference, consider the following pair of examples:

(4) a. I am not a Dortmund fan – I am a Schalke fan – but I am GLAD+ [Dortmund BEAT Bayern]$_{\text{TARGET}}$.

b. I am not a Dortmund fan – I am a Schalke fan – but I am GLAD Dortmund BEAT- [Bayern]$_{\text{TARGET}}$.

Example (4a) displays the stable, "literal" sentiment that is conveyed by the sentence: that the speaker is glad about the reported event. Example (4b), by contrast, displays an inferred sentiment: that the speaker specifically dislikes Bayern's team. The inferred sentiment toward Bayern may be canceled if the context was further elaborated, for instance by emphasizing a merely financial interest in the outcome ("If they hadn't, I would have lost my 100 € bet on that game").

Stoyanov and Cardie (2008) adopt a very pragmatic understanding of targets. They suggest a definition of opinion topic and present an algorithm for opinion topic identification that casts the task as a problem in topic co-reference resolution. In their work, they distinguish between:

"Topic The TOPIC of a fine-grained opinion is the real-world object, event or abstract entity that is the subject of the opinion as intended by the opinion source.

Topic span The TOPIC SPAN associated with an OPINION EXPRESSION is the closest, minimal span of text that mentions the topic.

Target span In contrast, TARGET SPAN denotes the span of text that covers the syntactic surface form comprising the contents of the opinion." (Stoyanov and Cardie, 2008, p. 818)

Notice the absence of any reference to syntactic relations between the subjective expression and the topic span, and the emphasis on the intentions of the opinion source for the identification of the topic. Given their definitions, Stoyanov and Cardie (2008) analyze the following example as indicated by the brackets and markup.

(5) [OH Al] THINKS that [TARGET SPAN [TOPIC SPAN? the government] should [TOPIC SPAN? tax gas] more in order to [TOPIC SPAN? curb [TOPIC SPAN? CO_2 emissions]]]. (= ex. (2), Stoyanov and Cardie, 2008, p. 818)

In example (5), the target span consists of the complement of *think* and there are multiple potential topics (denoted by the question marks in example 5) within the single target span of the opinion, each of them identified with its own topic span. This

illustrates that, at the text level, certain inferred targets might be more important than the overt target. In our annotations, targets correspond mostly to Stoyanov and Cardie (2008)'s target spans. What they consider as alternative topic spans relative to the same subjective expression is captured as targets of inferred opinions in our scheme and annotated in addition to the basic opinion that has their 'target span' as its target.

2.3 Prior Shared Tasks

While quite a few 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 (Strapparava and Mihalcea, 2007) *inter alia*), few tasks have included the extraction of sources and targets.

The most relevant prior work was done in the context of the Japanese NTCIR[3] 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 multiple opinions were expressed. The opinion source for a sentence could occur anywhere in the document. In the evaluation, where necessary, co-reference information was used to (manually) check whether a system response was part of the correct referent's chain of mentions. The sentences in the document were judged as either relevant (Y) or non-relevant (N) to the topic (=target). Polarity was determined for each opinionated sentence, and for sentences with more than one opinion expressed, the polarity of the main opinion expressed was chosen. All sentences were annotated by three assessors, allowing for strict and lenient (by majority vote) evaluation. The successor task, NTCIR-7: Multilingual Opinion Analysis (Seki et al., 2008), was basically similar in its setup to NTCIR-6, but also considered annotations relative to sub-sentences or clauses.

While the STEPS-task will focus on German, the most important difference to the shared tasks organized by NTCIR, as we will illustrate below, is that it defines the source and target extraction task at the level of individual subjective expressions. There is no shared task annotating at the expression level, rendering existing guidelines impractical and making the development of guidelines from scratch necessary. The corpus will be available for further annotation by ourselves and other research groups.

2.4 Corpora of political language

The usage of political corpora for NLP tasks is well-established within the scientific community. Thomas et al. (2006) collected US Congressional Speech Data, containing segments of uninterrupted speech. Guerini et al. (2008) constructed a corpus of tagged political speeches (CORPS), containing 3600 English-language speeches harvested from

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

the web. The authors focused on audience reactions and tagged applause or laughter to make these response signals usable as identifying markers of persuasive communications. Osenova and Simov (2012) built a corpus of Bulgarian political speeches containing both interviews with politicians as well as debates from the years 2006 to 2012. It has annotations for topic, turns, and linguistic units. Analysis of sentiment/opinions is in progress. Closer to our concerns in terms of the data used, Barbaresi (2012) constructed a corpus containing the political speeches by German presidents and chancellors (Bundespräsidentenkorpus: 1442 speeches (1984-2012); Bundeskanzlerkorpus: 1831 speeches (1998-2011)).

In a previous effort to create a gold-standard corpus for German opinion mining, IGGSA created MLSA, the Multi-Layered Sentiment Analysis corpus (Clematide et al., 2012). This corpus, consisting of 270 sentences crawled from news websites, is annotated at three levels: (i) the sentence-level, covering subjectivity and overall polarity of a sentence, (ii) word- and phrase-level, and (iii) expression-level, focusing on objective and direct speech events. While the expression-level annotation of the MLSA is similar in spirit to the annotations created here, the corpus as such is ultimately not suitable for our purposes because the sentences in the MLSA do not form full texts. They were sampled out of the larger Sdewac-Corpus (Faaß and Eckart (2013)), which contains parsable sentences from the web in scrambled order.

2.5 Summary

In summary, our annotation scheme picks up most of the linguistic features that have been pursued in related work. It is, however, ultimately distinct from prior work. For instance, we choose a simpler treatment in some cases such as targets where we follow grammar more closely and concentrate on arguments, whereas Stoyanov and Cardie (2008) are interested in topic spans with text-level relevance. In other cases, our treatment is implicit, as in the case of the nesting of sources, which, unlike Wiebe et al. (2005), we do not annotate explicitly. And, finally, unlike all prior shared tasks, we annotate at the expression level.

3 Definition of the STEPS-Shared Task

Given the difficulty of the tasks as well as the diversity of systems that researchers are working on, the STEPS shared task will offer one main task as well as two subtasks:

Main task Identification of subjective expressions with their respective sources and targets

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

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

We allow for participation in any combination of the tasks. However, so as to not give an unfair advantage to any participants, the main task is run and evaluated first before the gold information on subjective expressions is given out for the two subtasks, which will be run concurrently.

3.1 Data

The STEPS data set comes from the Swiss parliament (*Schweizer Bundesversammlung*). The choice of this particular data set is motivated as follows: (i) the source data is open to the public and allows for free distribution with the annotations[4]; (ii) the text allows for annotation of multiple sources and targets; (iii) the text meets the research interests of several IGGSA-members, i.e. supports collaborations with political scientists and researchers in digital humanities.

Since the Swiss parliament operates multi-lingually, we decided to discard not only non-German speeches but also German speeches that respond to, or comment on, speeches, heckling, and side questions in languages other than German. This was done so that no German data had to be annotated whose correct interpretation might depend on foreign-language material that our annotators might not be able to understand fully.

Additional potential difficulties derive from peculiarities of Swiss German found in the data. For instance, the vocabulary of Swiss German is different from standard German, often in subtle ways. For instance, the verb *vorprellen* is used in 6 instead of *vorpreschen*, which would be expected for German spoken in Germany.

(6) Es ist unglaublich: Weil die Aussenministerin vorgeprellt ist , kann man das nicht mehr zurücknehmen .
'It is incredible: because the foreign secretary acted rashly, we can't take that back again.'

In order to minimize any negative impact that might result from the misunderstanding of Swiss German by our German and Austrian annotators, we chose speeches related to what we considered non-parochial topics. For instance, we used texts related to international affairs rather than to Swiss municipal governance. In addition, the annotation guidelines encourage annotators to mark annotations as Swiss German when they involve language usage that they are not fully familiar with. Such cases can then be excluded or weighed differently for the purposes of system evaluation. In our annotation, such markings are in fact rare. We think this reflects the fact that although parliamentary speeches are medially spoken, they are conceptually written, and we find much less Swiss German vocabulary than one would expect in Swiss German colloquial speech (cf. Scherrer and Rambow (2010)).

The STEPS data set has the following pre-processing pipeline: sentence segmentation and tokenization using OpenNLP[5], lemmatization with the TreeTagger (Schmid, 1994), constituency parsing using the Berkeley parser (Petrov and Klein, 2007), and conversion

[4] We were not able to conclusively ascertain the copy rights for German parliamentary speeches.
[5] http://opennlp.apache.org/

	Exact Match	Partial Match
Subjective Expression	0.7634	0.8314
Sources (when SE match)	0.5685	0.5959
Targets (when SE match)	0.4521	0.7123

Table 1: Inter-annotator agreement for the second annotation step

of the parse trees into TigerXML-Format using TIGER-tools (Lezius, 2002). For the annotation we used the Salto-Tool (Burchardt et al., 2006).

3.2 Development of the annotation scheme

The different research interests of the IGGSA-members called for a novel annotation scheme, which we based on a first explorative annotation step. In this step, four annotators labeled a mutual set of 50 sentences with respect to opinions, targets and sources. The sole requirement was the annotation of sources and targets at the level of individual subjective expressions and consideration of all nested targets and holders. The annotators reported on annotation decisions to support the development of a first annotation scheme and formed an initial set of guidelines.

In a second step, two experienced annotators re-annotated the data using the initial guidelines and assessed them. The average inter-annotator agreement, i.e. the recall of annotations from both annotator perspectives, also took partial matches into consideration as proposed in Wiebe et al. (2005). Table 1 shows the results; we observed an agreement of 83% for subjective expressions (Wiebe et al. (2005) reports an average agreement of 72%) and 71% on targets. Cases of disagreement were subject to further analysis to enhance the guidelines.

3.3 Guidelines used

Generally, our annotation scheme can be characterized as a single-stage scheme aiming at full coverage.[6] That is, we only annotate at the expression level – we do not perform sentence or document-level annotations prior or subsequent to the expression-level annotation. And any and all kinds of subjective expressions by any source and on any topic were to be annotated. There was thus no focus on particular politicians, parties, issues etc. as potential sources or targets.

Our definition of subjective expressions is broad and based on well-known prototypes. It covers expressions of

- evaluation (positive or negative): *toll* 'great', *doof* 'stupid'
- (un)certainty: *zweifeln* 'doubt', *gewiss* 'certain'
- emphasis: *sicherlich/bestimmt* 'certainly'

[6] See `https://sites.google.com/site/iggsahome/downloads` for the final form of the guidelines.

- speech acts: *sagen* 'say', *ankündigen* 'announce'
- mental processes: *denken* 'think', *glauben* 'believe'

Our list of prototypes is inspired by, and largely overlaps with, the notions that Wiebe et al. (2005) subsumes under the umbrella term *private state*, following 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." However, beyond giving the prototypes, we did not seek to impose any particular theory from the linguistic or psychological literature related to subjectivity, appraisal, emotion or related notions.

We initially intended to distinguish polar facts from proper opinions. As we had conceived of the difference, polar facts were expressions whose status as subjective depended on context and for which even differences in polarity depending on context are conceivable (cf. 7a versus 7b), whereas real opinions result from the inherent meaning of words and syntactic patterns.

(7) The car interior uses a lot of plastic. (constructed)

a. That's great because it saves weight and, thus, gas.

b. It looks very cheap and inelegant.

However, we abandoned this plan after observing low agreement in intermediate rounds of annotation. In our final annotation round, polar facts could optionally be distinguished by setting a flag marking them as 'inferred' opinions on a subjective expression frame.

Further, no type of lexical or multi-word expression, or syntactic pattern was excluded from consideration. Thus, depending on the actual use in context, annotators could, for instance, mark as subjective expressions:

- exclamation marks
- rhetorical devices (marked also by a flag of the same name), chief among them:
  - repetitions (Ein Beschluss für Klimaschutz ist **an Deutschland gescheitert, an deutschen Abgeordneten, an Konservativen und Liberalen, ...** 'A proposal for climate protection failed because of Germany, because of German MPs, because of conservatives and liberals, ...' [7])
  - emphatically spelled words
  - rhetorical questions (**Und wer soll das bezahlen**? 'And who is supposed to pay for that?')

In identifying subjective expressions, annotators were instructed to select minimal spans where possible. This instruction went hand in hand with the decision that for the purposes of the shared task we would set aside any treatment of polarity and intensity.

[7] http://dip21.bundestag.de/dip21/btp/17/17240.pdf

Thus, negation, intensifiers and attenuators and any other expressions that might affect a minimal expression's polarity or intensity could be ignored.

An important aspect of the scheme is that the same expression could be labeled multiple times as a subjective expression with its own source and target. The need for this multi-layer annotation arises, for instance, in cases where a lexical item evokes two evaluations (cf. Maks and Vossen (2011)). The verb *prahlen* 'brag', for instance, conveys a positive evaluation by a participant in the event about another participant, and a second negative evaluation about an event participant by the speaker who uses the word *prahlen*. The need for multiple annotations also arises when multiple different semantic roles are evaluated. For instance, with verbs like *danken* 'thank' or *beschuldigen* 'accuse', arguably both a person and their behavior can be seen as targets of evaluation.

With respect to sources and targets, annotators were instructed to first consider syntactic/semantic dependents of the subjective expressions. If sources and targets were locally unexpressed, they could look further in the context and annotate other phrases.[8] In cases where a subjective expression represented the view of the implicit speaker/text author, annotators could set a flag 'Speaker' (*Sprecher*) on the source element. Note that the nesting of sources is not explicitly captured by our scheme. However, implicitly, it is captured as follows: a subjective expression A that is embedded within the target of another subjective expression B should have a source that is embedded under the source of expression B (see example (4) in Section 2.2).

4 Inter-annotator agreement

After the revision of the annotation guidelines as described above, five unseen speeches of the Swiss parliament, consisting of approximately 200 sentences, were selected for a proof-of-concept annotation round. Two groups, each consisting of three annotators, annotated about 100 sentences (two or three documents respectively). Both groups consisted of one experienced annotator and two master-level students, the latter having been trained for the annotations by a presentation of the annotation guidelines and example annotations. The inter-annotator agreement can be found in Tables 2 and 3. The first one shows the average pairwise inter-annotator agreement and the second one the agreement for the full-agreement mode, containing only those cases, where there was at least a partial match on the subjective expression level for all three annotators. All shown values include exact and partial matches. In addition, we always give the average dice coefficient (see equation 8), which we used for measuring the similarity of the annotations with respect to the overlapping terminals.

$$dice = \frac{2 * matching\ terminals}{terminals\ annotated\ by\ A1 + terminals\ annotated\ by\ A2} \tag{8}$$

[8] For the actual shared task, we plan on adding a layer of co-reference annotations to the data so that systems do not need to match a particular mention of the relevant source or target to receive credit.

	group 1			group 2		
	Armut1	**Aussen1**	**Aussen2**	**Buchpreis1**	**Buchpreis2**	**mean**[3]
Sources[1,2]	0.5375	0.4453	0.6742	0.7585	0.6605	0.6186
Dice across source matches	1.0000	0.9871	0.9977	0.9831	0.9896	0.9887
Targets[1,2]	0.6849	0.5384	0.5938	0.7883	0.6598	0.6549
Dice across target matches	0.7017	0.7058	0.7154	0.8406	0.8322	0.7722
Subjective Expression[1]	0.5728	0.4629	0.6456	0.5774	0.6554	0.5671
Dice across Subjective Expression matches[1]	0.8361	0.6538	0.5865	0.8901	0.7951	0.7563

[1] including exact and partial matches

[2] only considering cases with a match on the level of the subjective expression

[3] weighted by no. of sentences in the speeches

Table 2: Average pairwise inter-annotator agreement with a total number of annotated subjective expressions per annotator between 145 and 262 for group 1 and 122 and 236 for group 2

When comparing the agreement of the second annotation iteration (Table 1) and the proof-of-concept annotations (Table 2), a decrease in agreement of about 25%-points can be seen on the level of subjective expressions and a smaller decrease of about 6%-points to about 65.5% on the level of targets, but also a small increase of about 2%-points to 62% regarding source annotations. Considering that the annotators in the latter round were mostly unexperienced in this kind of task, and also considering that there were more annotators, leaving room for more disagreement, the results for the source and target annotations are quite satisfying, especially given the complexity of the annotation task. Compared to inter-annotator agreement studies of the previously mentioned NTCIR (M)OAT tasks, who reported an average pairwise agreement on opinionated judgements between $\kappa = 0.23$ (Chinese) and 0.67 (Japanese) in the first year (cf. Seki et al., 2007, p. 269) and 0.23 (English) and 0.71 (Japanese) in the second year (cf. Seki et al., 2008, p. 190, 193) and 0.46 (trad. Chinese) and 0.97 (simpl. Chinese) for the third year (cf. Seki et al., 2010, p. 214), the results are fairly good, bearing in mind, that the binary judgement of a complete sentence with respect to its opinionatedness is an easier task than actually identifying the subjective expression. Additionally, since the shared task primarily aims at addressing the challenge of identifying sources and targets of subjective expressions, the agreement on the subjective expressions themselves might be neglected. Nevertheless, we are going to closely examine the actual annotations in a qualitative error analysis and use the information gained thereby to further improve the annotation guidelines.

	group 1			group 2		
	Armut1	**Aussen1**	**Aussen2**	**Buchpreis1**	**Buchpreis2**	**mean**[2]
Source[1]	0.5000	0.3448	0.6552	0.6970	0.7429	0.5811
Target[1]	0.2000	0.2759	0.4483	0.7273	0.4000	0.4537
Subjective Expression	0.3155	0.2680	0.4987	0.4104	0.4829	0.3871

[1] only considering cases with a match on the level of the subjective expression
[2] weighted by no. of sentences

Table 3: Inter-annotator agreement for annotations with at least a partial match on the level of the subjective expression for all three annnotators (n=136)

5 Evaluation procedure

The runs that are submitted by the participants of the shared task, will be evaluated on different levels, according to the task they choose to participate in. For the full task, there will be 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 only the sources and targets will be evaluated, as the subjective expressions are already given.

The evaluation will be conducted in two different ways, based on the level of inter-annotator agreement in the gold standard annotations: The full-agreement mode will only consider annotations of the gold standard that have a match on the subjective expression level for all three annotators. The majority-vote mode uses the gold standard annotations where at least two of the three annotators agreed on the subjective expression level. We expect systems to perform better on the full-agreement subset, where human agreement is higher.

We use recall to measure the proportion of correct system annotations with respect to the gold standard annotations. Additionally, precision will be calculated to give the fraction of correct system annotations with respect to all the system annotations. For recall and precision in both modes of evaluation, we recognize a match when there is partial span overlap. Since full overlap on spans is relatively rare, we do not use a strict match criterion at all. Instead, we use the dice coefficient to measure the overlap between a system annotation and a gold standard annotation, in a way parallel to what we did for the measurement of inter-annotator agreement.

6 Conclusion

A complete understanding of opinions requires associating them with their sources and targets. While in some text types such as reviews the fillers of these roles can be readily guessed, they need to be retrieved from the actual text in many others. In order to allow for the evaluation of automatic systems on this complex task, we developed a shared task on the detection of targets, sources and subjective expressions. As our textual data, we selected political speeches from the Swiss parliament. They are particularly

suitable as they represent multiple topics, and contain multiple speakers and instances of nesting.

Using the guidelines that we developed through multiple rounds of annotation, we achieved reasonably high inter-annotator agreement. We also presented how we plan to evaluate the submissions of task participants. Our evaluation methods allow for a proper treatment of partial matches of annotation spans, and they distinguish cases of perfect agreement among annotators from cases which a majority but not all annotators labeled. The shared task will be held in the run-up of the KONVENS conference in 2014.

Acknowledgments

For their support in preparing and carrying out the annotations, we would like to thank Jasper Brandes, Melanie Dick, Inga Hannemann, and Daniela Schneevogt.

Literature

Barbaresi, A. (2012). German Political Speeches, Corpus and Visualization. Technical report, ENS Lyon. 2nd Version.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., and Pado, S. (2006). SALTO - A Versatile Multi-Level Annotation Tool. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 517–520.

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U., and Wiegand, M. (2012). Mlsa - a multi-layered reference corpus for german sentiment analysis. In Calzolari, N., Choukri, K., Declerck, T., Do?an, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Faaß, G. and Eckart, K. (2013). Sdewac – a corpus of parsable sentences from the web. In Gurevych, I., Biemann, C., and Zesch, T., editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg.

Guerini, M., Strapparava, C., and Stock, O. (2008). Corps: A corpus of tagged political speeches for persuasive communication processing. *Journal of Information Technology and Politics*, 5(1):19–32.

Lezius, W. (2002). TIGERsearch - Ein Suchwerkzeug für Baumbanken. In Busemann, S., editor, *Proceedings of KONVENS 2002*, Saarbrücken, Germany.

Maks, I. and Vossen, P. (2011). A verb lexicon model for deep sentiment analysis and opinion mining applications. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.0)*, pages 10–18, Portland, Oregon. Association for Computational Linguistics.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (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.

Osenova, P. and Simov, K. (2012). The Political Speech Corpus of Bulgarian. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Petrov, S. and Klein, D. (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. Association for Computational Linguistics.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Ruppenhofer, J., Somasundaran, S., and Wiebe, J. (2008). Finding the sources and targets of subjective expressions. In *LREC*, Marrakech, Morocco.

Scherrer, Y. and Rambow, O. (2010). Natural language processing for the swiss german dialect area. In Pinkal, M., Rehbein, I., Schulte im Walde, S., and Storrer, A., editors, *Semantic Approaches in Natural Language Processing*, pages 93–102.

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

Seki, Y., Evans, D., Ku, L.-W., Chen, H.-H., Kando, N., and Lin, C.-Y. (2007). Overview of opinion analysis pilot task at ntcir-6. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 265–278.

Seki, Y., Evans, D., Ku, L.-W., Sun, L., Chen, H.-H., Kando, N., and Lin, C.-Y. (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.

Seki, Y., Ku, L.-W., Sun, L., Chen, H.-H., and Kando, N. (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.

Stoyanov, V. and Cardie, C. (2008). Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 817–824, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.

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