Sentence and Expression Level Annotation of Opinions in User-Generated Discourse

Cigdem Toprak and Niklas Jakob and Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Computer Science Department, Technische Universität Darmstadt, Hochschulstraße 10
D-64289 Darmstadt, Germany
www.ukp.tu-darmstadt.de

Abstract

In this paper, we introduce a corpus of consumer reviews from the *rateitall* and the *eopinions* websites annotated with opinion-related information. We present a two-level annotation scheme. In the first stage, the reviews are analyzed at the sentence level for *(i)* relevancy to a given topic, and *(ii)* expressing an evaluation about the topic. In the second stage, on-topic sentences containing evaluations about the topic are further investigated at the expression level for pinpointing the properties (semantic orientation, intensity), and the functional components of the evaluations (opinion terms, targets and holders). We discuss the annotation scheme, the inter-annotator agreement for different subtasks and our observations.

1 Introduction

There has been a huge interest in the automatic identification and extraction of opinions from free text in recent years. Opinion mining spans a variety of subtasks including: creating opinion word lexicons (Esuli and Sebastiani, 2006; Ding et al., 2008), identifying opinion expressions (Riloff and Wiebe, 2003; Fahrni and Klenner, 2008), identifying polarities of opinions in context (Breck et al., 2007; Wilson et al., 2005), extracting opinion targets (Hu and Liu, 2004; Zhuang et al., 2006; Cheng and Xu, 2008) and opinion holders (Kim and Hovy, 2006; Choi et al., 2005).

Data-driven approaches for extracting opinion expressions, their holders and targets require reliably annotated data at the expression level. In previous research, expression level annotation of opinions was extensively investigated on newspaper articles (Wiebe et al., 2005; Wilson and Wiebe, 2005; Wilson, 2008b) and on meeting dialogs (Somasundaran et al., 2008; Wilson, 2008a).

Compared to the newspaper and meeting dialog genres, little corpus-based work has been carried out for interpreting the opinions and evaluations in user-generated discourse. Due to the high popularity of Web 2.0 communities[1], the amount of user-generated discourse and the interest in the analysis of such discourse has increased over the last years. To the best of our knowledge, there are two corpora of user-generated discourse which are annotated for opinion related information at the expression level: The corpus of Hu & Liu (2004) consists of customer reviews about consumer electronics, and the corpus of Zhuang et al. (2006) consists of movie reviews. Both corpora are tailored for application specific needs, therefore, do not contain certain related information explicitly annotated in the discourse, which we consider important (see Section 2). Furthermore, none of these works provide inter-annotator agreement studies.

Our goal is to create sentence and expression level annotated corpus of customer reviews which fulfills the following requirements: (1) It filters individual sentences regarding their topic relevancy and the existence of an opinion or factual information which implies an evaluation. (2) It identifies opinion expressions including the respective opinion target, opinion holder, modifiers, and anaphoric expressions if applicable. (3) The semantic orientation of the opinion expression is identified while considering negation, and the opinion expression is linked to the respective holder and target in the discourse. Such a resource would *(i)* enable novel applications of opinion mining such as a fine-grained identification of opinion properties, e.g. opinion modification detection including negation, and *(ii)* enhance opinion target extraction and the polarity assignment by linking the opinion expression with its target

[1] http://blog.nielsen.com/nielsenwire/wp-content/uploads/2008/10/press_release24.pdf

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 575–584,
Uppsala, Sweden, 11-16 July 2010.

and providing anaphoric resolutions in discourse.

We present an annotation scheme which fulfills the mentioned requirements, an inter-annotator agreement study, and discuss our observations. The rest of this paper is structured as follows: Section 2 presents the related work. In Sections 3, we describe the annotation scheme. Section 4 presents the data and the annotation study, while Section 5 summarizes the main conclusions.

2 Previous Opinion Annotated Corpora

2.1 Newspaper Articles and Meeting Dialogs

Most prominent work concerning the expression level annotation of opinions is the Multi-Perspective Question Answering (MPQA) corpus[2] (Wiebe et al., 2005). It was extended several times over the last years, either by adding new documents or annotating new types of opinion related information (Wilson and Wiebe, 2005; Stoyanov and Cardie, 2008; Wilson, 2008b). The MPQA annotation scheme builds upon the *private state* notion (Quirk et al., 1985) which describes mental states including opinions, emotions, speculations and beliefs among others. The annotation scheme strives to represent the *private states* in terms of their functional components (i.e. *experiencer* holding an *attitude* towards a *target*). It consists of frames (*direct subjective, expressive subjective element, objective speech event, agent, attitude,* and *target frames*) with slots representing various attributes and properties (e.g.*intensity, nested source*) of the private states.

Wilson (2008a) adapts and extends the concepts from the MPQA scheme to annotate subjective content in meetings (AMI corpus), and creates the AMIDA scheme. Besides subjective utterances, the AMIDA scheme contains *objective polar utterances* which annotates evaluations without expressing explicit opinion expressions.

Somasundaran et al. (2008) proposes *opinion frames* for representing discourse level associations in meeting dialogs. The annotation scheme focuses on two types of opinions, *sentiment* and *arguing*. It annotates the opinion expression and target spans. The *link* and *link type* attributes associate the target with other targets in the discourse through *same* or *alternative* relations. The *opinion frames* are built based on the links between targets. Somasundaran et al. (2008) show that *opinion frames* enable a coherent interpretation of the opinions in discourse and discover implicit evaluations through link transitivity.

Similar to Somasundaran et al. (2008), Asher et al. (2008) performs discourse level analysis of opinions. They propose a scheme which first identifies and assigns categories to the opinion segments as *reporting, judgment, advice,* or *sentiment*; and then links the opinion segments with each other via rhetorical relations including *contrast, correction, support, result*, or *continuation*. However, in contrast to our scheme and other schemes, instead of marking expression boundaries without any restriction they annotate an opinion segment only if it contains an opinion word from their lexicon, or if it has a rhetorical relation to another opinion segment.

2.2 User-generated Discourse

The two annotated corpora of user-generated content and their corresponding annotation schemes are far less complex. Hu & Liu (2004) present a dataset of customer reviews for consumer electronics crawled from amazon.com. The following example shows two annotations taken from the corpus of Hu & Liu (2004):

camera[+2]##This is my first digital camera and what a toy it is...

size[+2][u]##it is small enough to fit easily in a coat pocket or purse.

The corpus provides only target and polarity annotations, and do not contain opinion expression or opinion modifier annotations which lead to these polarity scores. The annotation scheme allows the annotation of implicit features (indicated with the the attribute *[u]*). Implicit features are not resolved to any actual product feature instances in discourse. In fact, the actual positions of the product features (or any anaphoric references to them) are not explicitly marked in the discourse, i.e, it is unclear to which mention of the feature the opinion refers to.

In their paper on movie review mining and summarization, Zhuang et al. (2006) introduce an annotated corpus of movie reviews from the Internet Movie Database. The corpus is annotated regarding movie features and corresponding opinions. The following example shows an annotated sentence:

⟨Sentence⟩I have never encountered a movie whose supporting cast was so perfectly realized.⟨FO Fword="supporting cast" Ftype="PAC" Oword="perfect" Otype="PRO"/⟩⟨/Sentence⟩

[2] http://www.cs.pitt.edu/mpqa/

The movie features (*Fword*) are attributed to one of 20 predefined categories (Ftype). The opinion words (*Oword*) and their semantic orientations (*Otype*) are identified. Possible negations are directly reflected by the semantic orientation, but not explicitly labeled in the sentence. *(PD)* in the following example indicates that the movie feature is referenced by anaphora:

⟨Sentence⟩It is utter nonsense and insulting to my intelligence and sense of history. ⟨FO Fword="film(PD)" Ftype="OA" Oword="nonsense, insulting" Otype="CON"/⟩⟨/Sentence⟩

However, similar to the corpus of Hu & Liu (2004) the referring pronouns are not explicitly marked in discourse. It is therefore neither possible to automatically determine which pronoun creates the link if there are more than one in a sentence, nor it is denoted which antecedent, i.e. the actual mention of the feature in the discourse it relates to.

3 Annotation Scheme

3.1 Opinion versus Polar Facts

The goal of the annotation scheme is to capture the evaluations regarding the topics being discussed in the consumer reviews. The evaluations in consumer reviews are either explicit expressions of opinions, or facts which imply evaluations as discussed below.

Explicit expressions of opinions: Opinions are private states (Wiebe et al., 2005; Quirk et al., 1985) which are not open to objective observation or verification. In this study, we focus on the opinions stating the quality or value of an entity, experience or a proposition from one's perspective. (1) illustrates an example of an explicit expression of an opinion. Similar to Wiebe et al. (2005), we view opinions in terms of their functional components, as *opinion holders*, e.g., the author in (1), holding attitudes (*polarity*), e.g., negative attitude indicated with the word *nightmare*, towards possible *targets*, e.g., *Capella University*.

(1) I had a nightmare with Capella University.[3]

Facts implying evaluations: Besides opinions, there are facts which can be objectively verified, but still imply an evaluation of the quality or value of an entity or a proposition. For instance, consider the snippet below:

[3]We use authentic examples from the corpus without correcting grammatical or spelling errors.

(2) In a 6-week class, I counted 3 comments from the professors directly to me and two directed to my team.
(3) I found that I spent most of my time learning from my fellow students.
(4) A standard response from my professors would be that of a sentence fragment.

The example above provides an evaluation about the professors without stating any explicit expressions of opinions. We call such objectively verifiable, but evaluative sentences *polar facts*. Explicit expressions of opinions typically contain specific cues, i.e. opinion words, loaded with a positive or negative connotation (e.g., *nightmare*). Even when they are taken out of the context in which they appear, they evoke an evaluation. However, evaluations in *polar facts* can only be inferred within the context of the review. For instance, the targets of the implied evalution in the polar facts (2), (3) and (4) are the professors. However, (3) may have been perceived as a positive statement if the review was explaining how good the fellow students were or how the course enforced team work etc.

The annotation scheme consists of two levels. First, the sentence level scheme analyses each sentence in terms of *(i)* its relevancy to the overall topic of the review, and *(ii)* whether it contains an evaluation (an opinion or a *polar fact*) about the topic. Once the on-topic sentences containing evaluations are identified, the expression level scheme first focuses either on marking the text spans of the opinion expressions (if the sentence contains an explicit expression of an opinion) or marking the targets of the *polar facts* (if the sentence is a *polar fact*). Upon marking an opinion expression span, the target and holder of the opinion is marked and linked to the marked opinion expression. Furthermore, the expression level scheme allows assigning polarities to the marked opinion expression spans and targets of the *polar facts*.

The following subsections introduce the sentence and the expression level annotation schemes in detail with examples.

3.2 Sentence Level Annotation

The sentence annotation strives to identify the sentences containing evaluations about the topic. In consumer reviews people occasionally drift off the actual topic being reviewed. For instance, as in (5) taken from a review about an online university, they tend to provide information about their background or other experiences.

(5) I am very fortunate and almost right out of high school

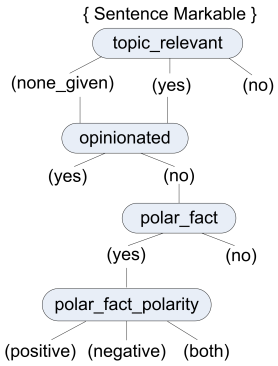

Figure 1: The sentence level annotation scheme

with a very average GPA and only 20; I already make above $45,000 a year as a programmer with a large health care company for over a year and have had 3 promotions up in the first year and a half.

Such sentences do not provide information about the actual topic, but typically serve for justifying the user's point of view or provide a better understanding about her circumstances. However, they are not valuable for an application aiming to extract opinions about a specific topic.

Reviews given to the annotators contain meta information stating the topic, for instance, the name of the university or the service being reviewed. A markable (i.e. an annotation unit) is created for each sentence prior to the annotation process. At this level, the annotation process is therefore a sentence labeling task. The annotators are able to see the whole review, and instructed to label sentences in the context of the whole review. Figure 1 presents the sentence level scheme. Attribute names are marked with oval circles and the possible values are given in parenthesis. The following attributes are used:

topic_relevant attribute is labeled as *yes* if the sentence discusses the given topic itself or its aspects, properties or features as in examples (1)-(4). Other possible values for this attribute include *none_given* which can be chosen in the absence of meta data, or *no* if the sentence drifted off the topic as in example (5).

opinionated attribute is labeled as *yes* if the sentence contains any explicit expressions of opinions about the given topic. This attribute is presented if the *topic_relevant* attribute has been labeled as *none_given* or *yes*. In other words, only the on-topic sentences are considered in this step. Examples (6)-(8) illustrate examples labeled as *topic_relevant=yes* and *opinionated=yes*.

*(6) Many people are knocking Devry but I have seen them to be a very great school. [**Topic:** Devry University]*
*(7) University of Phoenix was a surprising disappointment. [**Topic:** University of Phoenix]*
*(8) Assignments were passed down, but when asked to clarify the assignment because the syllabus had contradicting, poorly worded, information, my professors regularly responded...."refer to the syllabus"....but wait, the syllabus IS the question. [**Topic:** University of Phoenix]*

polar_fact attribute is labeled as *yes* if the sentence is a *polar fact*. This attribute is presented if the *opinionated* attribute has been labeled as *no*. Examples (2)-(4) demonstrate sentences labeled as *topic_relevant=yes, opinionated=no* and *polar_fact=yes*.

polar_fact_polarity attribute represents the polarity of the evaluation in a *polar_fact* sentence. The possible values for this attribute include *positive, negative, both*. The value *both* is intended for the *polar_fact* sentences containing more than one evaluation with contradicting polarities. At the expression level analysis, the targets of the contradicting *polar_fact* evaluations are identified distinctly and assigned polarities of *positive* or *negative* later on. Examples (9)-(11) demonstrate examples of *polar fact* sentences with different values of the attribute *polar_fact_polarity*.

*(9) There are students in the first programming class and after taking this class twice they cannot write a single line of code. **[polar_fact_polarity=negative]***
*(10) The same class (i.e. computer class) being teach at Ivy League schools are being offered at Devry. **[polar_fact_polarity=positive]***
*(11) The lectures are interactive and recorded, but you need a consent from the instructor each time. **[polar_fact_polarity=both]***

3.3 Expression Level Annotation

At the expression level, we focus on the topic relevant sentences containing evaluations, i.e., sentences labeled as *topic_relevant=yes, opinionated=yes* or *topic_relevant=yes, opinionated=no, polar_fact=yes*. If the sentence is a *polar fact*, then the aim is to mark the target and label the polarity of the evaluation. If the sentence is opinionated, then, the aim is to mark the opinion expression span, and label its polarity and strength (i.e. intensity), and to link it to the target and the holder.

Figure 2 presents the expression level scheme. At this stage, annotators mark text spans, and are allowed to assign one of the five labels to the marked span:

The ***polar_target*** is used to label the targets of the evaluations implied by *polar facts*. The *isReference* attribute labels *polar_targets* which are anaphoric references. The *polar_target_polarity*

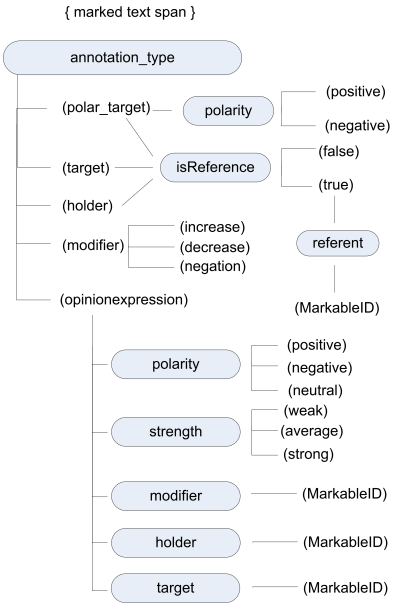

Figure 2: The expression level annotation scheme

attribute is used to label the polarity as *positive* or *negative*. If the *isReference* attribute is labeled as *true*, then the *referent* attribute appears which enables the annotator to resolve the reference to its antecedent. Consider the example sentences (12) and (13) below. The *polar_target* in (13), written bold, is labeled as *isReference=true*, *polar_target_polarity=negative*. To resolve the reference, annotator first creates another *polar_target* markable for the antecedent, namely the bold text span in (12), then, links the antecedent to the *referent* attribute of the *polar_target* in (13).

(12) Since classes already started, ***CTU*** *told me they would extend me so that I could complete the classes and get credit once I got back.*
(13) What ***they*** *didn't tell me is in order to extend, I also had to be enrolled in the next semester.*

The ***target*** annotation represents what the opinion is about. Both *polar_targets* and *targets* can be the topic of the review or different aspects, i.e. features of the topic. Similar to the *polar_targets*, the *isReference* attribute allows the identification of the targets which are anaphoric references and the *referent* attribute links them to their antecedents in the discourse. Bold span in (14) shows an example of a *target* in an opinionated sentence.

(14) Capella U has incredible ***faculty in the Harold Abel School of Psychology****.*

The ***holder*** type represents the holder of an opinion in the discourse and is labeled in the same manner as the *targets* and *polar_targets*. In consumer reviews, *holders* are most of the time the authors of the reviews. To ease the annotation process, the holder is not labeled when this is the author.

The ***modifier*** annotation labels the lexical items, such as *not, very, hardly etc.*, which affect the strength of an opinion or shift its polarity. Upon creation of a *modifier* markable, annotators are asked to choose between *negation, increase, decrease* for identifying the influence of the *modifier* on the opinion. For instance, the marked span in (15) is labeled as *modifier=increase* as it gives the impression that the author is really offended by the negative comments about her university.

(15) I am ***quite honestly*** *appauled by some of the negative comments given for Capella University on this website.*

The ***opinionexpression*** annotation is used to label the opinion terms in the sentence. This markable type has five attributes, three of which, i.e., *modifier*, *holder*, and *target* are pointer attributes to the previously defined markable types. The *polarity* attribute assesses the semantic orientation of the attitude, where the *strength* attribute marks the intensity of this attitude. The *polarity* and *strength* attributes focus solely on the marked *opinionexpression* span, not the whole evaluation implied in the sentence. For instance, the *opinionexpression* span in (16) is labeled as *polarity=negative, strength=average*. We infer the polarity of the evaluation only after considering the *modifier, polarity* and the *strength* attributes together. In (16), the evaluation about the target is strongly negative after considering all three attributes of the *opinionexpression* annotation. In (17), the *polarity* of the *opinionexpression1* itself (*complaints*) is labeled as *negative*. It is linked to the *modifier1* which is labeled as *negation*. *Target1* (*PhD journey*) is linked to the *opinionexpression1*. The overall evaluation regarding the *target1* is positive after applying the affect of the *modifier1* to the *polarity* of the *opinionexpression1*, i.e., after negating the negative polarity.

(16) I am ***quite honestly*** $_{[modifier]}$ ***appauled by*** $_{[opinionexpression]}$ *some of the* ***negative comments given for Capella University on this website*** $_{[target]}$.
(17) I have ***no*** $_{[modifier1]}$ ***complaints*** $_{[opinionexpression1]}$ *about the entire* ***PhD journey*** $_{[target1]}$ *and* ***highly*** $_{[modifier2]}$ ***recommend*** $_{[opinionexpression2]}$ ***this school*** $_{[target2]}$.

Finally, Figure 3 demonstrates all expression level markables created for an opinionated sentence and how they relate to each other.

Figure 3: Expression level annotation example

4 Annotation Study

Each review has been annotated by two annotators independently according to the annotation scheme introduced above. We used the freely available MMAX2[4] annotation tool capable of stand-off multi-level annotations. Annotators were native speaker linguistic students. They were trained on 15 reviews after reading the annotation manual.[5] In the training stage, the annotators discussed with each other if different decisions have been made and were allowed to ask questions to clarify their understanding of the scheme. Annotators had access to the review text as a whole while making their decisions.

4.1 Data

The corpus consists of consumer reviews collected from the review portals *rateitall*[6] and *epinions*[7]. It contains reviews from two domains including online universities, e.g., Capella University, Pheonix, University of Maryland University College etc. and online services, e.g., PayPal, egroups, eTrade, eCircles etc. These two domains were selected with the project-relevant, domain-specific research goals in mind. We selected a specific topic, e.g. Pheonix, if there were more than 3 reviews written about it. Table 1 shows descriptive statistics regarding the data.

We used 118 reviews containing 1151 sentences from the university domain for measuring the sentence and expression level agreements. In the following subsections, we report the inter-annotator agreement (IAA) at each level.

[4] http://mmax2.sourceforge.net/

[5] http://www.ukp.tu-darmstadt.de/research/data/sentiment-analysis

[6] http://www.rateitall.com

[7] http://www.epinions.com

	University	Service	All
Reviews	240	234	474
Sentences	2786	6091	8877
Words	49624	102676	152300
Avg sent./rev.	11.6	26	18.7
Std. dev. sent./rev.	8.2	16	14.6
Avg. words/rev.	206.7	438.7	321.3
Std. dev. words/rev.	159.2	232.1	229.8

Table 1: Descriptive statistics about the corpus

4.2 Sentence Level Agreement

Sentence level markables were already created automatically prior to the annotation, i.e., the set of annotation units were the same for both annotators. We use Cohen's kappa (κ) (Cohen, 1960) for measuring the IAA. The sentence level annotation scheme has a hierarchical structure. A new attribute is presented based on the decision made for the previous attribute, for instance, *opinionated* attribute is only presented if the *topic_relevant* attribute is labeled as *yes* or *none_given*; *polar_fact* attribute is only presented if the *opinionated* attribute is labeled as *no* etc. We calculate κ for each attribute considering only the markables which were labeled the same by both annotators in the previously required step. Table 2 shows the κ values for each attribute, the size of the markable set on which the value was calculated, and the percentage agreement.

Attribute	Markables	Agr.	κ
topic_relevant	1151	0.89	0.73
opinionated	682	0.80	0.61
polar_fact	258	0.77	0.56
polar_fact_polarity	103	0.96	0.92

Table 2: Sentence level inter-annotator agreement

The agreement for topic relevancy shows that it is possible to label this attribute reliably. The sentences labeled as topic relevant by both annotators correspond to 59% of all sentences, suggesting that people often drift off the topic in consumer reviews. This is usually the case when they provide information about their backgrounds or alternatives to the given topic.

On the other hand, we obtain moderate agreement levels for the *opinionated* and *polar_fact* attributes. 62% of the topic relevant sentences were labeled as opinionated by at least one annotator, and the rest 38% constitute the topic relevant sentences labeled as not opinionated by both annotators. Nonetheless, they still contain evaluations (*polar_facts*), as 15% of the topic relevant sen-

tences were labeled as *polar_facts* by both annotators. When we merge the attributes *opinionated* and *polar_fact* into a single category, we obtain κ of 0.75 and a percentage agreement of 87%. Thus, we conclude that opinion-relevant sentences, either in the form of an explicit expression of opinion or a *polar fact*, can be labeled reliably in consumer reviews. However, there is a thin border between *polar facts* and explicit expressions of opinions.

To the best of our knowledge, similar annotation efforts on consumer or movie reviews do not provide any agreement figures for direct comparison. However, Wiebe et al. (2005) present an annotation study where they mark textual spans for subjective expressions in a newspaper corpus. They report pairwise κ values for three annotators ranging between 0.72 - 0.84 for the sentence level subjective/objective judgments. Wiebe et al. (2005) mark subjective spans, and do not explicitly perform the sentence level labeling task. They calculate the sentence level κ values based on the existence of a subjective expression span in the sentence. Although the task definitions, approaches and the corpora have quite disparate characteristics in both studies, we obtain comparable results when we merge *opinionated* and *polar_fact* categories.

4.3 Expression Level Agreement

At the expression level, annotators focus only on the sentences which were labeled as *opinionated* or *polar_fact* by both annotators. Annotators were instructed to mark text spans, and then, assign them the annotation types such as *polar_target, opinionexpression* etc. (see Figure 2). For calculating the text span agreement, we use the agreement metric presented by Wiebe et al. (2005) and Somasundaran et al. (2008). This metric corresponds to the precision (P) and recall (R) metrics in information retrieval where the decisions of one annotator are treated as the system; the decisions of the other annotator are treated as the gold standard; and the overlapping spans correspond to the correctly retrieved documents.

Somasundaran et al. (2008) present a discourse level annotation study in which opinion and target spans are marked and linked with each other in a meeting transcript corpus. Following Somasundaran et al. (2008), we compute three different measures for the text span agreement: *(i) exact matching* in which the text spans should perfectly match; *(ii) lenient (relaxed) matching* in which the overlap between spans is considered as a match, and *(iii) subset matching* in which a span has to be contained in another span in order to be considered as a match.[8] Agreement naturally increases as we relax the matching constraints. However, there were no differences between the *lenient* and the *subset* agreement values. Therefore, we report only the *exact* and *lenient matching* agreement results for each annotation type in Table 3. The same agreement results for the *lenient* and *subset* matching indicates that inexact matches are still very similar to each other, i.e., at least one span is totally contained in the other.

Somasundaran et al. (2008) do not report any F-measure. However, they report span agreement results in terms of precision and recall ranging between 0.44 - 0.87 for opinion spans and between 0.74 - 0.90 for the target spans. Wiebe et al. (2005) use the *lenient matching* approach for reporting text span agreements ranging between 0.59 - 0.81 for subjective expressions. We obtain higher agreement values for both opinion expression and target spans. We attribute this to the fact that the annotators look for opinion expression and target spans within the opinionated sentences which they agreed upon. Sentence level analysis indeed increases the reliability at the expression level. Compared to the high agreement on marking *target* spans, we obtain lower agreement values on marking *polar_target* spans. We observe that it is easier to attribute explicit expressions of evaluations to topic relevant entities compared to attributing evaluations implied by experiences to specific topic relevant entities in the reviews.

We calculated the agreement on identifying anaphoric references using the method introduced in (Passonneau, 2004) which utilizes Krippendorf's α (Krippendorff, 2004) for computing reliability for coreference annotation. We considered the overlapping *target* and *polar_target* spans together in this calculation, and obtained an α value of 0.29. Compared to Passonneau (α values from 0.46 to 0.74), we obtain a much lower agreement value. This may be due to the different definitions and organizations of the annotation tasks. Passonneau requires prior marking of all noun phrases (or instances which needs to be processed by the an-

[8] An example of *subset matching*: waste of time vs. total waste of time

Span	Exact			Lenient		
	P	R	F	P	R	F
opinionexpression	0.70	0.80	0.75	0.82	0.93	0.87
modifier	0.80	0.82	0.81	0.86	0.86	0.86
target	0.80	0.81	0.80	0.91	0.90	0.91
holder	0.75	0.72	0.73	0.93	0.88	0.91
polar_target	0.67	0.42	0.51	0.75	0.49	0.59

Table 3: Inter-annotator agreement on text spans at the expression level

notator). Annotator's task is to identify whether an instance refers to another marked entity in the discourse, and then, to identify corefering entity chains. However, in our annotation process annotators were tasked to identify only one entity as the referent, and was free to choose it from anywhere in the discourse. In other words, our chains contain only one entity. It is possible that both annotators performed correct resolutions, but still did not overlap with each other, as they resolve to different instances of the same entity in the discourse. We plan to further investigate reference resolution annotation discrepancies and perform corrections in the future.

Some annotation types require additional attributes to be labeled after marking the span. For instance, upon marking a text span as a *polar_target* or an *opinionexpression*, one has to label the *polarity* and *strength*. We consider the overlapping spans for each annotation type and use κ for reporting the agreement on these attributes. Table 4 shows the κ values.

Attribute	Markables	Agr.	κ
polarity	329	0.97	0.94
strength	329	0.74	0.55
modifier	136	0.88	0.77
polar_target_polarity	63	0.80	0.67

Table 4: Inter-annotator agreement at the expression level

We observe that the *strength* of the *opinionexpression* and the *polar_target_polarity* cannot be labeled as reliably as the *polarity* of the *opinionexpression*. 61% of the agreed upon *polar_targets* were labeled as negative by both annotators. On the other hand, only 35% of the agreed upon *opinionexpressions* were labeled as negative by both annotators. There were no neutral instances. This indicates that reviewers tend to report negative experiences using polar facts, probably objectively describing what has happened, but report positive experiences with explicit opinion expressions. Distribution of the *strength* attribute was as follows: *weak* 6%, *average* 54%, and *strong* 40%. The majority of the modifiers were annotated as intensifiers (70%), while 20% of the modifiers were labeled as negation.

4.4 Discussion

We analyzed the discrepancies in the annotations to gain insights about the challenges involved in various opinion related labeling tasks. At the sentence level, there were several trivial cases of disagreement, for instance, failing to recognize topic relevancy when the topic was not mentioned or referenced explicitly in the sentence, as in (18). Occasionally, annotators disagreed about whether a sentence that was written as a reaction to the other reviewers, as in (19), should be considered as topic relevant or not. Another source of disagreement included sentences similar to (20) and (21). One annotator interpreted them as universally true statements regardless of the topic, while the other attributed them to the discussed topic.

(18) Go to a state university if you know whats good for you!
(19) Those with sour grapes couldnt cut it, have an ax to grind, and are devoting their time to smearing the school.
(20) As far as learning, you really have to WANT to learn the material.
(21) On an aside, this type of education is not for the undisciplined learner.

Annotators easily distinguished the evaluations at the sentence level. However, they had difficulties distinguishing between a *polar_fact* and an opinion. For instance, both annotators agreed that the sentences (22) and (23) contain evaluations regarding the topic of the review. However, one annotator interpreted both sentences as objectively verifiable facts giving a positive impression about the school, while the other one treated them as opinions.

(22) All this work in the first 2 Years!
(23) The school has a reputation for making students work really hard.

Sentence level annotation increases the reliability of the expression level annotation in terms of marking text spans. However, annotators often had disagreements on labeling the *strength* attribute. For instance, one annotator labeled the

opinion expression in (24) as *strong*, while the other one labeled it as *average*. We observe that it is not easy to identify trivial causes of disagreements regarding *strength* as its perception by each individual is highly subjective. However, most of the disagreements occurred between *weak* and *average* cases.

(24) the experience that i have when i visit student finance is much ***like going to the dentist****, except when i leave, nothing is ever fixed.*

We did not apply any consolidation steps during our agreement studies. However, a final version of the corpus will be produced by the third judge (one of the co-authors) by consolidating the judgements of the two annotators.

5 Conclusions

We presented a corpus of consumer reviews from the *rateitall* and *eopinions* websites annotated with opinion related information. Existing opinion annotated user-generated corpora suffer from several limitations which result in difficulties for interpreting the experimental results and for performing error analysis. To name a few, they do not explicitly link the functional components of the opinions like targets, holders, or modifiers with the opinion expression; some of them do not mark opinion expression spans, none of them resolves anaphoric references in discourse. Therefore, we introduced a two level annotation scheme consisting of the sentence and expression levels, which overcomes the limitations of the existing review corpora. The sentence level annotation labels sentences for *(i)* relevancy to a given topic, and *(ii)* expressing an evaluation about the topic. Similar to (Wilson, 2008a), our annotation scheme allows capturing evaluations made with factual (objective) sentences. The expression level annotation further investigates on-topic sentences containing evaluations for pinpointing the properties (polarity, strength), and marking the functional components of the evaluations (opinion terms, modifiers, targets and holders), and linking them within a discourse. We applied the annotation scheme to the consumer review genre and presented an extensive inter-annotator study providing insights to the challenges involved in various opinion related labeling tasks in consumer reviews. Similar to the MPQA scheme, which is successfully applied to the newspaper genre, the annotation scheme treats opinions and evaluations as a composition of functional components and it is easily extendable. Therefore, we hypothesize that the scheme can also be applied to other genres with minor extensions or as it is. Finally, the corpus and the annotation manual will be made available at `http://www.ukp.tu-darmstadt.de/research/data/sentiment-analysis`.

Acknowledgements

This research was funded partially by the German Federal Ministry of Economy and Technology under grant 01MQ07012 and partially by the German Research Foundation (DFG) as part of the *Research Training Group on Feedback Based Quality Management in eLearning* under grant 1223. We are very grateful to Sandra Kübler for her help in organizing the annotators, and to Lizhen Qu for his programming support in harvesting the data.

References

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2008. Distilling opinion in discourse: A preliminary study. In *Coling 2008: Companion volume: Posters*, pages 7–10, Manchester, UK.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-2007)*, pages 2683–2688, Hyderabad, India.

Xiwen Cheng and Feiyu Xu. 2008. Fine-grained opinion topic and polarity identification. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2710–2714, Marrekech, Morocco.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362, Morristown, NJ, USA.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008*, pages 231–240, Palo Alto, California, USA.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 417–422, Genova, Italy.

Angela Fahrni and Manfred Klenner. 2008. Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proceedings of the Symposium on Affective Language in Human and Machine, AISB 2008 Convention*, pages 60 – 63, Aberdeen, Scotland.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD'04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, Washington.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text at the joint COLING-ACL Conference*, pages 1–8, Sydney, Australia.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methology*. Sage Publications, Thousand Oaks, California.

Rebecca J. Passonneau. 2004. Computing reliability for coreference. In *Proceedings of LREC*, volume 4, pages 1503–1506, Lisbon.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, New York.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *EMNLP-03: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 105–112.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2008. Discourse level opinion relations: An annotation study. In *In Proceedings of SIGdial Workshop on Discourse and Dialogue*, pages 129–137, Columbus, Ohio.

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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

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.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada.

Theresa Wilson. 2008a. Annotating subjective content in meetings. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Theresa Ann Wilson. 2008b. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States.* Ph.D. thesis, University of Pittsburgh.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50, Arlington, Virginia, USA.