Annotating Topics of Opinions

Veselin Stoyanov, Claire Cardie

Department of Computer Science, Cornell University
Ithaca, NY 14850, USA
ves@cornell.edu, cardie@cornell.edu

Abstract

Fine-grained subjectivity analysis has been the subject of much recent research attention. As a result, the field has gained a number of working definitions, technical approaches and manually annotated corpora that cover many facets of subjectivity. Little work has been done, however, on one aspect of fine-grained opinions – the specification and identification of opinion topics. In particular, due to the difficulty of manual opinion topic annotation, no general-purpose opinion corpus with information about topics of fine-grained opinions currently exists. In this paper, we propose a methodology for the manual annotation of opinion topics and use it to annotate a portion of an existing general-purpose opinion corpus with opinion topic information. Inter-annotator agreement results according to a number of metrics suggest that the annotations are reliable.

1. Introduction

Subjectivity analysis is concerned with extracting information about any attitudes, beliefs, emotions, opinions, evaluations and sentiment expressed in texts. The field has received much recent attention, motivated at least in part by the wide range of information analysis applications that research in the area can support (see, e.g. Coglianese (2004) and Stoyanov et al. (2005)) and by the challenging problems in computational linguistics and natural language learning that subjectivity analysis engenders.

In contrast to *coarse-grained subjectivity analysis*, which is concerned with identifying subjectivity or sentiment at the document level (e.g. Pang et al. (2002), Turney (2002a)), we are interested in *fine-grained subjectivity analysis* — the identification, extraction and characterization of subjective language at the phrase or clause level.

The area of fine-grained subjectivity analysis has seen multiple research efforts, resulting in several definitions of what constitutes an expression of opinion and what components make up each expression of opinion (see Related Work, Section 3.). Although variations exist, researchers almost universally agree that an expression of opinion is characterized by its source, its polarity and its topic or target. In addition, research in the area has been further facilitated by the creation of several corpora that have been manually annotated with fine-grained expressions of opinions, including their source and polarity (Bethard et al., 2004; Wiebe et al., 2005). Notably missing from the corpora, however, are annotations for the topics of opinions. Despite the desire and motivation for creating such corpora, topic annotation has proven a difficult task (Wilson, 2005; Wiebe, 2005). Nonetheless, topics remain an important component of an opinion, and topic extraction remains a critical step for sentiment analysis systems.

In this paper, we describe a methodology for performing opinion topic annotation. We apply the methodology to extend an existing opinion corpus with topic information. We discuss our progress on corpus creation and present the results of an inter-annotator agreement study using several reliability measures. Our results indicate that opinion topic annotation is a feasible task using the new annotation methodology.

2. Terminology

As mentioned above, although precise definitions of what constitutes an opinion sometimes differ, a general specification of the components that characterize fine-grained expressions of opinion has emerged. We will illustrate these components, or attributes, using the following two examples.

(1)John likes Prof. Smith for his upbeat attitude.

(2)John believes that there will be a question about Malaria on the midterm.

An opinion is characterized by the following components:

- **Opinion Expression.** The span of text signaling the expression of an opinion — the underlined words "likes" in Example (1) and "believes" in Example (2).
- **Source.** The opinions in both examples can be attributed to "John", who is the *source* or *opinion-holder* (shown in bold).
- **Polarity.** The opinion in Example (1) expresses John's positive feelings, so it is said that it has a positive *polarity*. The opinion in the second example does not carry a specific sentiment, so it is said to exhibit a neutral *polarity*.

In addition, we use the following definitions in our discussion of the fourth component of an opinion, the opinion *topic*:

- **Topic.** The real-world object, event, or abstract entity that is the primary subject of the opinion as intended by the opinion holder. The topic of the opinion in Example (1) is the person PROF. SMITH. In Example (2), the topic is not as clear. We could argue that it is either THE MIDTERM or MALARIA (and will discuss this in more detail in Section 4.).

- **Topic Span.** The *topic span* associated with an opinion expression is the closest minimal span of text that mentions the topic. In both examples, the topic span is the text that designates the topic entity ("Prof. Smith" and "midterm", or "Malaria", respectively).
- **Target Span.** In contrast, we use *target span* to denote the span of text that covers the syntactic surface form comprising the contents of the opinion. In Example (1), the topic span and target span coincide, while in Example (2), the target span consists of the complement to the opinion expression verb — the text "there will be a question about Malaria on the midterm."

3. Related Work

3.1. Sentiment Analysis

Research in sentiment analysis can be divided into two major categories based on the granularity of the opinions involved — coarse-grained sentiment analysis, concerned with opinions at the document level, and fine-grained sentiment analysis, concerned with opinion recognition at the clause or phrase level or below. Our research falls in the latter category.

The problem of sentiment extraction at the document level (*sentiment classification*) has been tackled as a text categorization task in which the goal is to assign to a document either positive ("thumbs up") or negative ("thumbs down") polarity (e.g. Das and Chen (2001), Pang et al. (2002), Turney (2002b), Dave et al. (2003), Pang and Lee (2004)).

Recent work in the area of fine-grained opinion analysis, has offered several different definitions of what constitutes an expression of opinion. For example, Bethard et al. (2004) define an opinion as a sentence or part of a sentence that would answer the question "What does X feel about Y?", while Wiebe et al. (2005) center their definition around Quirk et al.'s (1985) notion of a private state, defined as "a state that is not open to objective observation or verification."

Using their definition of opinion, Wiebe et al. (2005) have created an opinion annotation scheme covering *subjective expressions* — any expression of a private state in text. They further apply their annotation scheme to create the MPQA corpus[1], which consists of 535 documents manually annotated for phrase-level expressions of opinions, their sources and polarities.[2] Other efforts have attempted to create resources for fine-grained opinion analysis (e.g. Voorhees and Buckland (2003), Bethard et al. (2004)), but we are not aware of a corpus that rivals the scale and depth of the MPQA corpus.

It should be noted that Wiebe et al. (2005) initially intended to include topic annotations in the MPQA corpus, but postponed the task, discovering that topic annotation was very difficult (Wilson, 2005; Wiebe, 2005). Currently, Wiebe et al. are adding *target spans* to their annotations. While useful, target spans are insufficient for many applications that use fine-grained opinion information: they neither contain information indicating which opinions are about the same topic, nor provide a concise representation of the topic.

[1] Available at www.cs.pitt.edu/mpqa/databaserelease.

[2] The strength, or intensity, of the opinion is also annotated.

Creation of the language resources described above has encouraged work on automatically extracting different aspects of opinions. Recent work has used the MPQA corpus (e.g. Riloff and Wiebe (2003), Wilson et al. (2004), Wiebe and Riloff (2005), Choi et al. (2005)) as well as other resources (e.g. Dave et al. (2003), Bethard et al. (2004), Yu and Hatzivassiloglou (2003)) to show that systems can be trained to recognize opinion expressions, their sources, their polarity and their strength at reasonable levels of accuracy.

3.2. Opinion Topic Identification

In the domain of product reviews, several research efforts have tackled the extraction of the topic of the opinion (e.g. Yi et al. (2003), Hu and Liu (2004), Kobayashi et al. (2004), Popescu and Etzioni (2005)). In this genre of text, however, it has been adequate to limit the notion of topic to mentions of product names and components and their attributes. Thus, topic extraction has been effectively substituted with a lexicon look-up and techniques have focused on how to learn or acquire an appropriate lexicon for the task.

Because the existing general opinion corpora do not contain sufficient information on opinion topics, the problem of opinion topic extraction has been largely unexplored in NLP. A notable exception is the work of Kim and Hovy (2006), who propose a model that extracts opinion topics for subjective expressions signaled by verbs and adjectives. Their model relies on semantic frames and extracts as the topic the syntactic constituent at a specific argument position for the given verb or adjective. In other words, Kim and Hovy extract what we refer to as the target spans, and do so for a subset of the opinion-bearing words in the text. Although on many occasions target spans coincide with opinion topics (as in Example (1)), we have observed that on many other occasions this is not the case (as in Example (2)). Furthermore, hampered by the lack of resources with manually annotated targets, Kim and Hovy could provide only a limited evaluation.

4. The Topic Annotation Methodology

As discussed above, existing work on the annotation of opinion topics focuses on target spans. As an example, consider the following sentence:

(3)President Chen Shui-bian has on many occasions expressed goodwill toward *mainland China*.

In some cases, as in (3), opinion topics are expressed clearly as the single noun phrase that comprises the target span. Existing approaches for topic annotation can easily deal with these cases. In general, we have observed that opinion topics are realized as target spans more frequently when the opinion carries sentiment as in (3). This leads us to conjecture that opinion topic identification should be easier for sentiment-bearing opinions.

(4)"It all depends on how mainland China interprets President Chen's latest remarks on cross-strait relations and how the two

sides cultivate an environment favorable for resumption of their long-stalled dialogue," **Tsai** explained.

Example (4) illustrates the difficulty of opinion topic annotation. The sentence clearly contains an opinion signaled by the predicate "explained". However, it is very hard to pinpoint a single phrase that states the topic — there are multiple potential such phrases depending on the intended topic of the opinion: "mainland China", "President Chen's latest remarks", "cross-strait relations", "resumption of their long-stalled dialogue", or even "President Chen" himself. In general, and in cases like these in particular, we argue that the topic of the opinion depends critically on the context. Consider example (5) below:

(5)Tsai Ing-wen said Tuesday she foresees the possibility of the two sides of the Taiwan Strait resuming dialogue next year.

If sentence (4) immediately follows sentence (5), we would argue that, based on the context, the topic of the opinion in (4) is referred to via the topic span phrase "the resumption of the cross-straight dialogue". As a result, we consider a sensible strategy for topic annotation to be the following: **assign to the opinion the topic that constitutes the primary information goal of the opinion expression**. For example, if sentence (4) were in an article talking about President Chen's political career, we could argue that the topic of opinion (4) is PRESIDENT CHEN himself, denoted via the topic span *President Chen*.

This context-dependent definition allows us to deal with the difficulty of pinpointing the topic of an opinion. However, we have introduced new problems. First, a topic might not be mentioned as a simple noun phrase within the opinion sentence. Moreover, due to the context, the annotator may wish to change his or her definition of the topic over the course of the document. For instance, in the case where PRESIDENT CHEN is the topic, the annotator might realize that the document is about President Chen only after reading a good part of the document and annotating several opinions. The new difficulties of topic annotation stem from the fact that an annotator has to remember the set of topics introduced in the document and judge whether a new opinion should be considered coreferent with one of the existing topics or should indicate the onset of a new topic.

To deal with these difficulties, we introduce the notion of *topic-coreferent opinions*. We consider two opinions to be topic-coreferent if they are about the same general topic. For example, the opinion from example (4) is topic-coreferent with the opinion from example (5) if (4) follows (5) in the document.

Using our new definition, we perform topic annotation by

- identifying all topic-coreferent opinions, and then
- labeling each cluster of topic-coreferent opinions with a descriptive string.

We believe that judging topic coreference will be a relatively easy task (both for people and computers) because annotators can take advantage of contextual and positional clues.

4.1. Opinion Topic Annotation Process

Our topic annotation process begins with a corpus annotated w.r.t. fine-grained expressions of opinions. (For the current work we use the MPQA corpus.) To facilitate the opinion annotation process we developed a set of annotation instructions based on the preceding discussion[3] and a graphical user interface (GUI) that helps the annotator to keep track of the existing topics. Aided by the GUI, an annotator proceeds as follows:

1. The annotator opens a manually annotated opinion document. The GUI shows three panels (i) a panel containing a list of all opinions that are yet to be annotated — initially all opinions in the document (where each opinion is characterized by the words that signal the expression of the opinion, its source and its polarity), (ii) an initially empty panel that contains the current set of topic-coreferent clusters and, (iii) a panel containing the text of the document.

2. The annotator proceeds down the list of opinions that are yet to be annotated. Looking at the clusters of topic-coreferent opinions in panel (ii) as well as the text in panel (iii), the annotator decides whether the current opinion is coreferent with the opinions in any of the existing clusters or should start a new topic. The annotator then drags the opinion to the appropriate cluster in panel (ii).

3. After dropping all opinions into the appropriate cluster, the annotator assigns a label to name each cluster, based on the opinions in the cluster[4].

4. In addition, we require the annotator to mark the spans of text that contributed to the topic coreference decision, since learning algorithms may benefit from this information. More specifically, the annotator marks the topic spans, which we view as secondary information, but that can still be important for training automatic opinion identifiers. We allow the annotator to mark the topic spans at any time during the annotation process.

5. Finally, the annotator saves the document. The GUI checks the annotations to make sure that all opinions are assigned to a topic cluster, that all clusters are labeled and that all opinions are assigned a topic span.

5. Inter-annotator Agreement Study

We selected for annotation a random subset of 150 documents containing two or more opinions from the 535 documents in the MPQA corpus. Of these, we selected at random a subset of 20 documents to be annotated by two annotators for the purpose of performing an inter-annotator agreement study.

[3] Available at www.cs.cornell.edu/~ves.

[4] In reality, the annotator may assign a label to a cluster before assigning all opinions in the document. Indeed, we encourage the annotator to maintain a working label for each cluster.

5.1. Evaluation Metrics

The heart of our approach is the topic coreference judgment for opinions. For the inter-annotator agreement study, we compare these judgments across the pair of annotation sets (one for each annotator) and evaluate them via several measures borrowed from studies of noun phrase coreference resolution. We present these metrics in the next subsections.

5.1.1. Krippendorff's α

As one evaluation measure, we use Passonneau's (2004) generalization of Krippendorff's α (Krippendorff, 1980) — a standard metric employed for inter-annotator reliability studies. Krippendorff's α is a theoretically-founded measure with a nice probabilistic interpretation. It is designed to measure the reliability of coding agreement. Passonneau's innovation makes it possible to apply the α statistic to coreference clusters. Unfortunately, in its new formulation the measure does not carry the original probabilistic interpretation.

5.1.2. MUC score

The MUC score is a model-theoretic coreference scoring metric for noun phrase coreference resolution (Vilain et al., 1995). The MUC recall score is computed as the ratio of correct non-repetitive links in the response (i.e. the system's output) as compared to the minimum number of non-repetitive links required to construct the key (i.e. the gold standard topic clusters). The MUC precision score is computed by reversing the roles of the key and the response.

The MUC score has proved an intuitive and useful coreference resolution metric. However, a number of flaws in the scoring algorithm have been identified. Importantly, the algorithm does not credit responses that correctly include singleton clusters (i.e. clusters that contain only one item) and, in general, it is not strict enough for for responses that link too many clusters together.

5.1.3. B-Cubed

B-Cubed (B^3) is another commonly used noun phrase coreference resolution evaluation measure (Bagga and Baldwin, 1998). It is computed as the precision and recall for each item (in our case, each opinion) and is then averaged for each document. The precision (recall) for an item i is computed as the proportion of items in the intersection of the response and key clusters containing the item divided by the number of items in the response (key) cluster.

5.1.4. CEAF

As an example of another group of coreference measures that rely on mapping response clusters to key (gold standard) clusters, we selected Luo's (2005) CEAF score (short for Constrained Entity-Alignment F-Measure). The CEAF score is similar to a simplified version of the ACE score (ACE, 2005). It works by computing an optimal mapping of response clusters to key clusters, summing the scores for each pair of mapped clusters and dividing by the maximum score (i.e. the score for mapping the key to itself).

	α	B^3	CEAF	MUC
All opinions	.5476	.6424	.6904	.8383
Sentiment ops.	.7285	.7180	.7967	.8068
Strong ops.	.7669	.7374	.8217	.8209

Table 1: Inter-annotator agreement results.

	α	B^3	CEAF	MUC
One cluster	-.1017	.3739	.2976	.9200
One per cluster	.2238	.2941	.2741	-
Same paragraph	.3123	.5542	.5090	.7932

Table 2: Baseline results.

5.2. Results and Discussion

Results for inter-annotator agreement were computed on the 20 documents annotated by two annotators and are presented in Table 1. The table shows the overall agreement for all opinions (in the first row). Additionally, we argued intuitively, and observed from the data, that it is easier to label topics for opinions carrying sentiment. Our observations are confirmed empirically as annotator agreement improves when we evaluate it over only the sentiment-bearing opinions (row two) and over the strongly sentiment-based opinions[5] (row three).

A problem with using coreference resolution scoring algorithms for our inter-annotator agreement studies is that it is hard to translate absolute scores to quality of agreement. Of the four metrics, only Krippendorff's α attempts to incorporate a probabilistic interpretation (Passonneau, 2004). It is generally agreed that an α score above 0.66 indicates reliable agreement. Our inter-annotator agreement exhibits a score under that threshold when computed over all opinions (0.54) and a score above the threshold when computed over the sentiment-bearing opinions (0.71). However, as discussed above, in adapting α to the problem of coreference resolution, the score loses its probabilistic interpretation. For example, the α score requires that a pairwise distance function between clusters is specified. We used one sensible choice for such a function (we measured the distance between clusters A and B as $dist(A, B) = (2 * |A \cap B|)/(|A| + |B|)$), but other sensible choices for the distance lead to much higher scores.

Clearly, the numerical magnitudes of the inter-annotator agreement scores are insufficient to judge the quality of the annotation agreement. To be able to compare meaningfully the inter-annotator agreement scores, we compute scores comparing the clusters for one of the annotators to three "baselines" shown in Table 2, with the purpose of approximating chance agreement. The first baseline clusters the opinions by putting all of the opinions from a document into a single cluster (results shown in the first row of the table). The second baseline puts each opinion in its own cluster (shown in the second row), while the third baseline forms a cluster for all opinions that come from the same

[5]These are identified using the manually annotated strength, i.e. intensity, values.

paragraph in the document (shown in the third row). As Table 2 shows, all baselines score significantly lower than the inter-annotator agreement with the exception of the MUC score, which we discuss below. Furthermore, the baseline that groups opinions by paragraph appears to agree much better with the annotator, which is to be expected given our understanding of the way that topics in general, and opinion topics in particular, are expressed in discourse. This result leads us to believe that opinion topic annotation can be performed reliably.

Additionally, Table 2 points to a problem in using the MUC score for opinion topic coreference. Namely, topic coreference clusters tend to be much larger than noun phrase coreference clusters. This means that there are very few "non-links" to be recognized. The MUC score has some well-documented problems in not being strict enough for punishing clusterings that fail to identify "non-links" (Bagga and Baldwin, 1998). As a result, when comparing two opinion topic clusterings, it is very difficult to score better than the simple baseline of putting all opinions in the same cluster, which achieves perfect recall and a high precision (in our experiment, the precision was 0.838 for the MUC F-score of 0.920 in Table 2).

6. Conclusions

We presented a new methodology for opinion topic annotation that is based on the notion of opinion topic coreference. Using our new methodology and a relatively simple GUI, we annotated 150 of the 535 documents in the MPQA corpus w.r.t. opinion topics. An inter-annotator agreement study performed on 20 documents annotated by two annotators indicates that the annotations are reliable.

7. Acknowledgements

We would like to thank Janyce Wiebe and Theresa Wilson for helpful discussion and Benjamin Cole, a Cornell undergraduate in Information Science, for performing most of the annotation. This work was supported in part by National Science Foundation Grants BCS-0624277 and IIS-0535099 and by Department of Homeland Security Grant N0014-07-1-0152.

8. References

ACE. 2005. The NIST ACE evaluation website. http://www.nist.gov/speech/tests/ace/.

A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *In Proceedings of MUC7*.

S. Bethard, H. Yu, A. Thornton, V. Hativassiloglou, and D. Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of EMNLP*.

C. Coglianese. 2004. E-rulemaking: Information technology and regulatory policy: New directions in digital government research. Technical report, Harvard University, J. F. Kennedy School of Government.

S. Das and M. Chen. 2001. Yahoo for amazon: Extracting market sentiment from stock message boards. In *Proceedings of APFAAC*.

K. Dave, S. Lawrence, and D. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of IWWWC*.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, pages 755–760.

S. Kim and E. Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text*.

N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of IJCNLP*.

K. Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

X. Luo. 2005. On coreference resolution performance metrics. In *Proceedings of EMNLP*.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.

R. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of LREC*.

A. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*.

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

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceesings of EMNLP*.

V. Stoyanov, C. Cardie, and J. Wiebe. 2005. Multi-Perspective question answering using the OpQA corpus. In *Proceedings of EMNLP*.

P. Turney. 2002a. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*.

P. Turney. 2002b. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of MUC6*.

E. Voorhees and L. Buckland. 2003. Overview of the TREC 2003 Question Answering Track. In *Proceedings of TREC 12*.

J. Wiebe and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing*.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

Janyce Wiebe. 2005. Personal communications.

T. Wilson, J. Wiebe, and R. Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI*.

Theresa Wilson. 2005. Personal communications.

J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of ICDM*.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*.