Topic Identification for Fine-Grained Opinion Analysis

Veselin Stoyanov and **Claire Cardie**
Department of Computer Science
Cornell University
{stoyanov,cardie}@cs.cornell.edu

Abstract

Within the area of general-purpose fine-grained subjectivity analysis, *opinion topic identification* has, to date, received little attention due to both the difficulty of the task and the lack of appropriately annotated resources. In this paper, we provide an operational definition of *opinion topic* and present an algorithm for opinion topic identification that, following our new definition, treats the task as a problem in topic coreference resolution. We develop a methodology for the manual annotation of opinion topics and use it to annotate topic information for a portion of an existing general-purpose opinion corpus. In experiments using the corpus, our topic identification approach statistically significantly outperforms several non-trivial baselines according to three evaluation measures.

1 Introduction

Subjectivity analysis is concerned with extracting information about attitudes, beliefs, emotions, opinions, evaluations, sentiment and other private states expressed in texts. In contrast to the problem of identifying subjectivity or sentiment at the document level (e.g. Pang et al. (2002), Turney (2002)), we are interested in *fine-grained subjectivity analysis*, which is concerned with subjectivity at the phrase or clause level. We expect fine-grained subjectivity analysis to be useful for question-answering, summarization, information extraction and search engine support for queries of the form "How/what does entity X feel/think about topic Y?", for which document-level opinion analysis methods can be problematic.

Fine-grained subjectivity analyses typically identify SUBJECTIVE EXPRESSIONS in context, characterize their POLARITY (e.g. positive, neutral or negative) and INTENSITY (e.g. weak, medium, strong, extreme), and identify the associated SOURCE, or OPINION HOLDER, as well as the TOPIC, or TARGET, of the opinion. While substantial progress has been made in automating some of these tasks, opinion topic identification has received by far the least attention due to both the difficulty of the task and the lack of appropriately annotated resources.[1]

This paper addresses the problem of topic identification for fine-grained opinion analysis of general text.[2] We begin by providing a new, operational definition of *opinion topic* in which the topic of an opinion depends on the context in which its associated opinion expression occurs. We also present a novel method for general-purpose opinion topic identification that, following our new definition, treats the problem as an exercise in topic coreference resolution. We evaluate the approach using the existing MPQA corpus (Wiebe et al., 2005), which we extend with manual annotations that encode topic information (and refer to hereafter as the $\text{MPQA}_{\text{TOPIC}}$ corpus).

Inter-annotator agreement results for the manual annotations are reasonably strong across a number of metrics and the results of experiments that evaluate our topic identification method in the context of fine-grained opinion analysis are promising:

[1] Section 3 on related work provides additional discussion.

[2] The identification of products and their components and attributes from product reviews is a related, but quite different task from that addressed here. Section 3 briefly discusses, and provides references, to the most relevant research in that area.

using either automatically or manually identified topic spans, we achieve topic coreference scores that statistically significantly outperform two topic segmentation baselines across three coreference resolution evaluation measures (B^3, α and CEAF). For the B^3 metric, for example, the best baseline achieves a topic coreference score on the $\text{MPQA}_{\text{TOPIC}}$ corpus of 0.55 while our topic coreference algorithm scores 0.57 and 0.71 using automatically, and manually, identified topic spans, respectively.

In the remainder of the paper, we define opinion topics (Section 2), present related work (Section 3), and motivate and describe the key idea of topic coreference that underlies our methodology for both the manual and automatic annotation of opinion topics (Section 4). Creation of the $\text{MPQA}_{\text{TOPIC}}$ corpus is described in Section 5 and our topic identification algorithm, in Section 6. The evaluation methodology and results are presented in Sections 7 and 8, respectively.

2 Definitions and Examples

Consider the following opinion sentences:

(1)[OH John] adores [TARGET+TOPIC SPAN Marseille] and visits it often.

(2)[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]]].

A fine-grained subjectivity analysis should identify: the OPINION EXPRESSION[3] as "adores" in Example 1 and "thinks" in Example 2; the POLARITY as positive in Example 1 and neutral in Example 2; the INTENSITY as medium and low, respectively; and the OPINION HOLDER (OH) as "John" and "Al", respectively. To be able to discuss the opinion TOPIC in each example, we begin with three definitions:

– **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 holder.

– **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, 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, for instance, "Marseille" is both the TOPIC SPAN and the TARGET SPAN associated with the city of Marseille, which is the TOPIC of the opinion. In Example 2, the TARGET SPAN consists of the text that comprises the complement of the subjective verb "thinks". Example 2 illustrates why opinion topic identification is difficult: within the single target span of the opinion, there are multiple potential topics, each identified with its own topic span. Without more context, however, it is impossible to know which phrase indicates the intended topic. If followed by sentence 3, however,

(3) Although he doesn't like government-imposed taxes, he thinks that a fuel tax is the only effective solution.

the topic of Al's opinion in 2 is much clearer — it is likely to be fuel tax, denoted via the TOPIC SPAN "tax gas" or "tax".

3 Related Work

As previously mentioned, there has been much recent progress in extracting fine-grained subjectivity information from general text. Previous efforts have focused on the extraction of opinion expressions in context (e.g. Bethard et al. (2004), Breck et al. (2007)), the assignment of polarity to these expressions (e.g. Wilson et al. (2005), Kim and Hovy (2006)), source extraction (e.g. Bethard et al. (2004), Choi et al. (2005)), and identification of the source-expresses-opinion relation (e.g. Choi et al. (2006)), i.e. linking sources to the opinions that they express.

Not surprisingly, progress has been driven by the creation of language resources. In this regard, Wiebe et al.'s (2005) opinion annotation scheme for *subjective expressions* was used to create the MPQA corpus, which consists of 535 documents manually annotated for phrase-level expressions of opinions, their sources, polarities, and intensities. Although other opinion corpora exist (e.g. Bethard et al. (2004), Voorhees and Buckland (2003), the product review corpora of Liu[4]), we are not aware of any corpus that rivals the scale and depth of the MPQA corpus.

In the related area of opinion extraction from product reviews, several research efforts have focused on the extraction of the topic of the opinion (e.g. Kobayashi et al. (2004), Yi et al. (2003),

[3] For simplicity, we will use the term *opinion* throughout the paper to cover all types of private states expressed in subjective language.

[4] http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

Popescu and Etzioni (2005), Hu and Liu (2004)). For this specialized text genre, it has been sufficient 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. While the techniques have been very successful for this genre of text, they have not been applied outside the product reviews domain. Further, there are analyses (Wiebe et al., 2005) and experiments (Wilson et al., 2005) that indicate that lexicon-lookup approaches to subjectivity analysis will have limited success on general texts.

Outside the product review domain, there has been little effort devoted to opinion topic annotation. The MPQA corpus, for example, was originally intended to include topic annotations, but the task was abandoned after confirming that it was very difficult (Wiebe, 2005; Wilson, 2005), although target span annotation is currently underway. While useful, target spans alone will be insufficient for many applications: they neither contain information indicating which opinions are about the same topic, nor provide a concise textual representation of the topics.

Due to the lack of appropriately annotated corpora, the problem of opinion topic extraction has been largely unexplored in NLP. A notable exception is the work of Kim and Hovy (2006). They 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.

As we have defined it, opinion topic identification bears some resemblance to topic segmentation, the goal of which is to partition a text into a linear sequence of topically coherent segments. Existing methods for topic segmentation typically assume that fragments of text (e.g. sentences or sequences of words of a fixed length) with similar lexical distribution are about the same topic; the goal of these methods is to find the boundaries where the lexical distribution changes (e.g. Choi (2000), Malioutov and Barzilay (2006)). Opinion topic identification differs from topic segmentation in that opinion topics are not necessarily spatially coherent — there may be two opinions in the same sentence on different topics, as well as opinions that are on the same topic separated by opinions that do not share that topic. Nevertheless, we will compare our topic identification approach to a state-of-the-art topic segmentation algorithm (Choi, 2000) in the evaluation.

Other work has successfully adopted the use of clustering to discover entity relations by identifying entities that appear in the same sentence and clustering the intervening context (e.g. Hasegawa et al. (2004), Rosenfeld and Feldman (2007)). This work, however, considers named entities and heads of proper noun phrases rather than topic spans, and the relations learned are those commonly held between NPs (e.g. senator-of-state, city-of-state, chairman-of-organization) rather than a more general coreference relation.

4 A Coreference Approach to Topic Identification

Given our initial definition of opinion topics (Section 2), the next task is to determine which computational approaches might be employed for automatic opinion topic identification. We begin this exercise by considering some of the problematic characteristics of opinion topics.

Multiple potential topics. As noted earlier via Example 2, a serious problem in opinion topic identification is the mention of multiple potential topics within the target span of the opinion. Although an issue for all opinions, this problem is typically more pronounced in opinions that do not carry sentiment (as in Example 2). Our current definition of opinion topic requires the NLP system (or a human annotator) to decide which of the entities described in the target span, if any, refers to the intended topic. This decision can be aided by the following change to our definition of opinion topic, which introduces the idea of a context-dependent information focus: *the* TOPIC *of an opinion is the real-world entity that is the subject of the opinion as intended by the opinion holder* **based**

on the discourse context.

With this modified definition in hand, and given Example 3 as the succeeding context for Example 2, we argue that the intended subject, and hence the TOPIC, of Al's opinion in 2 can be quickly identified as the FUEL TAX, which is denoted by the TOPIC SPANS "tax gas" in 2 and "fuel tax" in 3.

Opinion topics not always explicitly mentioned. In stark contrast to the above, on many occasions the topic is not mentioned explicitly at all within the target span, as in the following example:

(5)[$_{\text{OH}}$ John] identified the violation of Palestinian human rights as one of the main factors. TOPIC: ISRAELI-PALESTINIAN CONFLICT

We have further observed that the opinion topic is often not mentioned within the same paragraph and, on a few occasions, not even within the same document as the opinion expression.

4.1 Our Solution: Topic Coreference

With the above examples and problems in mind, we hypothesize that the notion of *topic coreference* will facilitate both the manual and automatic identification of opinion topics: **We say that two opinions are topic-coreferent if they share the same opinion topic.** In particular, we conjecture that judging whether or not two opinions are topic-coreferent is easier than specifying the topic of each opinion (due to the problems described above).

5 Constructing the MPQA$_{\text{TOPIC}}$ Corpus

Relying on the notion of topic coreference, we next introduce a new methodology for the manual annotation of opinion topics in text:

1. The annotator begins with a corpus of documents that has been annotated w.r.t. OPINION EXPRESSIONS. With each opinion expression, the corpus provides POLARITY and OPINION HOLDER information. (We use the aforementioned MPQA corpus.)

2. The annotator maintains a list of the opinion expressions that remain to be annotated (initially, all opinion expressions in the document) as well as a list of the current groupings (i.e. clusters) of opinion expressions that have been identified as topic-coreferent (initially this list is empty).

3. For each opinion expression, in turn, the annotator decides whether the opinion is on the same topic as the opinions in one of the existing clusters or should start a new cluster, and inserts the opinion in the appropriate cluster.

4. The annotator labels each cluster with a string that describes the opinion topic that covers all opinions in the cluster.

5. The annotator marks the TOPIC SPAN of each opinion. (This can be done at any point in the process.)

The manual annotation procedure is described in a set of instructions available at http://www.cs.cornell.edu/~ves. In addition, we created a GUI that facilitates the annotation procedure. With the help of these resources, one person annotated opinion topics for a randomly selected set of 150 of the 535 documents in the MPQA corpus. In addition, 20 of the 150 documents were selected at random and annotated by a second annotator for the purposes of an inter-annotator agreement study, the results of which are presented in Section 8.1. The MPQA$_{\text{TOPIC}}$ and the procedure by which it was created are described in more detail in (Stoyanov and Cardie, 2008).

6 The Topic Coreference Algorithm

As mentioned in Section 4, our computational approach to opinion topic identification is based on topic coreference: For each document (1) find the clusters of coreferent opinions, and (2) label the clusters with the name of the topic. In this paper we focus only on the first task, topic coreference resolution — the most critical step for topic identification. We conjecture that the second step can be performed through frequency analysis of the terms in each of the clusters and leave it for future work.

Topic coreference resolution resembles another well-known problem in NLP — noun phrase (NP) coreference resolution. Therefore, we adapt a standard machine learning-based approach to NP coreference resolution (Soon et al., 2001; Ng and Cardie, 2002) for our purposes. Our adaptation has three steps: (i) identify the topic spans; (ii) perform pairwise classification of the associated opinions as to whether or not they are topic-coreferent; and, (iii) cluster the opinions according to the results of (ii). Each step is discussed in more detail below.

6.1 Identifying Topic Spans

Decisions about topic coreference should depend on the text spans that express the topic. Ideally, we would be able to recover the topic span of each opinion and use its content for the topic coreference decision. However, the topic span depends on the topic itself, so it is unrealistic that topic spans can be recovered with simple methods. Nevertheless, in this initial work, we investigate two simple methods for automatic topic span identification and compare them to two manual approaches:

- **Sentence.** Assume that the topic span is the whole sentence containing the opinion.
- **Automatic.** A rule-based method for identifying the topic span (developed using MPQA documents that are not part of $\text{MPQA}_{\text{TOPIC}}$). Rules depend on the syntactic constituent type of the opinion expression and rely on syntactic parsing and grammatical role labeling.
- **Manual.** Use the topic span marked by the human annotator. We included this method to provide an upper bound on performance of the topic span extractor.
- **Modified Manual.** Meant to be a more realistic use of the manual topic span annotations, this method returns the manually identified topic span only when it is within the sentence of the opinion expression. When this span is outside the sentence boundary, this method returns the opinion sentence.

Of the 4976 opinions annotated across the 150 documents of $\text{MPQA}_{\text{TOPIC}}$, the topic spans associated with 4293 were within the same sentence as the opinion; 3653 were within the span extracted by our topic span extractor. Additionally, the topic spans of 173 opinions were outside of the paragraph containing the opinion.

6.2 Pairwise Topic Coreference Classification

The heart of our method is a pairwise topic coreference classifier. Given a pair of opinions (and their associated polarity and opinion holder information), the goal of the classifier is to determine whether the opinions are topic-coreferent. We use the manually annotated data to automatically learn the pairwise classifier. Given a training document, we construct a training example for every pair of opinions in the document (each pair is represented as a feature vector). The pair is labeled as a positive example if the two opinions belong to the same topic cluster, and a negative example otherwise.

Pairwise coreference classification relies critically on the expressiveness of the features used to describe the opinion pair. We use three categories of features: positional, lexico-semantic and opinion-based features.

Positional features These features are intended to exploit the fact that opinions that are close to each other are more likely to be on the same topic. We use six positional features:

- **Same Sentence/Paragraph**[5] True if the two opinions are in the same sentence/paragraph.
- **Consecutive Sentences/Paragraphs** True if the two opinions are in consecutive sentences/paragraphs.
- **Number of Sentences/Paragraphs** The number of sentences/paragraphs that separate the two opinions.

TOPIC SPAN**-based lexico-semantic features** The features in this group rely on the topic spans and are recomputed w.r.t. each of the four topic span methods. The intuition behind this group of features is that topic-coreferent opinions are likely to exhibit lexical and semantic similarity within the topic span.

- **tf.idf** The cosine similarity of the tf.idf weighted vectors of the terms contained in the two spans.
- **Word overlap** True if the two topic spans contain any contain words in common.
- **NP coref** True if the two spans contain NPs that are determined to be coreferent by a simple rule-based coreference system.
- **NE overlap** True if the two topic spans contain named entities that can be considered aliases of each other.

Opinion features The features in this group depend on the attributes of the opinion. In the current work, we obtain these features directly from the manual annotations of the $\text{MPQA}_{\text{TOPIC}}$ corpus, but they might also be obtained from automatically identified opinion information using the methods referenced in Section 3.

- **Source Match** True if the two opinions have the same opinion holder.
- **Polarity Match** True if the two opinions have the same polarity.

[5] We use sentence/paragraph to describe two features – one based on the sentence and one on the paragraph.

- **Source-Polarity Match** False if the two opinions have the same opinion holder but conflicting polarities (since it is unlikely that a source will have two opinions with conflicting polarities on the same topic).

We employ three classifiers for pairwise coreference classification – an averaged perceptron (Freund and Schapire, 1998), SVM^{light} (Joachims, 1998) and a rule-learner – RIPPER (Cohen, 1995). However, we report results only for the averaged perceptron, which exhibited the best performance.

6.3 Clustering

Pairwise classification provides an estimate of the likelihood that two opinions are topic-coreferent. To form the topic clusters, we follow the pairwise classification with a clustering step. We selected a simple clustering algorithm – single-link clustering, which has shown good performance for NP coreference. Given a threshold, single-link clustering proceeds by assigning pairs of opinions with a topic-coreference score above the threshold to the same topic cluster and then performs transitive closure of the clusters.[6]

7 Evaluation Methodology

For training and evaluation we use the 150-document $\text{MPQA}_{\text{TOPIC}}$ corpus. All machine learning methods were tested via 10-fold cross validation. In each round of cross validation, we use eight of the data partitions for training and one for parameter estimation (we varied the threshold for the clustering algorithm), and test on the remaining partition. We report results for the three evaluation measures of Section 7 using the four topic span extraction methods introduced in Section 6. The threshold is tuned separately for each evaluation measure. As noted earlier, all runs obtain opinion information from the $\text{MPQA}_{\text{TOPIC}}$ corpus (i.e. this work does not incorporate automatic opinion extraction).

7.1 Topic Coreference Baselines

We compare our topic coreference system to four baselines. The first two are the "default" baselines:

- **one topic** – assigns all opinions to the same cluster.
- **one opinion per cluster** – assigns each opinion to its own cluster.

The other two baselines attempt to perform topic segmentation (discussed in Section 3) and assign all opinions within the same segment to the same opinion topic:

- **same paragraph** – simple topic segmentation by splitting documents into segments at paragraph boundaries.
- **Choi 2000** – Choi's (2000) state-of-the-art approach to finding segment boundaries. We use the freely available C99 software described in Choi (2000), varying a parameter that allows us to control the average number of sentences per segment and reporting the best result on the test data.

7.2 Evaluation Metrics

Because there is disagreement among researchers w.r.t. the proper evaluation measure for NP coreference resolution, we use three generally accepted metrics[7] to evaluate our topic coreference system.

B-CUBED. B-CUBED (B^3) is a commonly used NP coreference metric (Bagga and Baldwin, 1998). It calculates precision and recall for each item (in our case, each opinion) based on the number of correctly identified coreference links, and then computes the average of the item scores in each document. Precision/recall for an item i is computed as the proportion of items in the intersection of the response (system-generated) and key (gold standard) clusters containing i divided by the number of items in the response/key cluster.

CEAF. As a representative of another group of coreference measures that rely on mapping response clusters to key clusters, we selected Luo's (2005) CEAF score (short for Constrained Entity-Alignment F-Measure). Similar to the ACE (2005) score, CEAF operates by computing an optimal mapping of response clusters to key clusters and assessing the goodness of the match of each of the mapped clusters.

Krippendorff's α. Finally, we use Passonneau's (2004) generalization of Krippendorff's (1980) α — a standard metric employed for inter-annotator

[6] Experiments using best-first and last-first clustering approaches provided similar or worse results.

[7] The MUC scoring algorithm (Vilain et al., 1995) was omitted because it led to an unjustifiably high MUC F-score (.920) for the ONE TOPIC baseline.

	B^3	α	CEAF
All opinions	.6424	.5476	.6904
Sentiment opinions	.7180	.7285	.7967
Strong opinions	.7374	.7669	.8217

Table 1: Inter-annotator agreement results.

reliability studies. Krippendorff's α is based on a probabilistic interpretation of the agreement of coders as compared to agreement by chance. While Passonneau's innovation makes it possible to apply Krippendorff's α to coreference clusters, the probabilistic interpretation of the statistic is unfortunately lost.

8 Results

8.1 Inter-annotator Agreement

As mentioned previously, out of the 150 annotated documents, 20 were annotated by two annotators for the purpose of studying the agreement between coders. Inter-annotator agreement results are shown in Table 1. We compute agreement for three subsets of opinions: all available opinions, only the sentiment-bearing opinions and the subset of sentiment-bearing opinions judged to have polarity of medium or higher.

The results support our conjecture that topics of sentiment-bearing opinions are much easier to identify: inter-annotator agreement for opinions with non-neutral polarity (SENTIMENT OPINIONS) improves by a large margin for all measures. As in other work in subjectivity annotation, we find that strong sentiment-bearing opinions are easier to annotate than sentiment-bearing opinions in general.

Generally, the α score aims to probabilistically capture the agreement of annotation data and separate it from chance agreement. It is generally accepted that an α score of .667 indicates reliable agreement. The score that we observed for the overall agreement was an α of .547, which is below the generally accepted level, while α for the two subsets of sentiment-bearing opinions is above .72. However, as discussed above, due to the way that it is adapted to the problem of coreference resolution, the α 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. Furthermore, we observed that the behavior of the α score can be rather erratic — small changes in one of the clusterings can lead to big differences in the score.

	B^3	α	CEAF
One topic	.3739	-.1017	.2976
One opinion per cluster	.2941	.2238	.2741
Same paragraph	.5542	.3123	.5090
Choi	.5399	.3734	.5370
Sentence	.5749	.4032	.5393
Rule-based	.5730	.4056	.5420
Modified manual	.6416	.5134	.6124
Manual	.7097	.6585	.6184

Table 2: Results for the topic coreference algorithms.

Perhaps a better indicator of the reliability of the coreference annotation is a comparison with the baselines, shown in the top half of Table 2. All baselines score significantly lower than the inter-annotator agreement scores. With one exception, the inter-annotator agreement scores are also higher than those for the learning-based approach (results shown in the lower half of Table 2), as would typically be expected. The exception is the classifier that uses the manual topic spans, but as we argued earlier these spans carry significant information about the decision of the annotator.

8.2 Baselines

Results for the four baselines are shown in the first four rows of Table 2. As expected, the two baselines performing topic segmentation show substantially better scores than the two "default" baselines.

8.3 Learning methods

Results for the learning-based approaches are shown in the bottom half of Table 2. First, we see that each of the learning-based methods outperforms the baselines. This is the case even when sentences are employed as a coarse substitute for the true topic span. A Wilcoxon Signed-Rank test shows that differences from the baselines for the learning-based runs are statistically significant for the B^3 and α measures ($p < 0.01$); for CEAF, using sentences as topic spans for the learning algorithm outperforms the SAME PARAGRAPH baseline ($p < 0.05$), but the results are inconclusive when compared with the system of CHOI.

In addition, relying on manual topic span information (MANUAL and MODIFIED MANUAL) allows the learning-based approach to perform significantly better than the two runs that use automatically identified spans ($p < 0.01$, for all three measures). The improvement in the scores hints at the importance of improving automatic topic span extraction, which will be a focus of our future work.

9 Conclusions

We presented a new, operational definition of opinion topics in the context of fine-grained subjectivity analysis. Based on this definition, we introduced an approach to opinion topic identification that relies on the identification of topic-coreferent opinions. We further employed the opinion topic definition for the manual annotation of opinion topics to create the $\text{MPQA}_{\text{TOPIC}}$ corpus. Inter-annotator agreement results show that opinion topic annotation can be performed reliably. Finally, we proposed an automatic approach for identifying topic-coreferent opinions, which significantly outperforms all baselines across three coreference evaluation metrics.

Acknowledgments The authors of this paper would like to thank Janyce Wiebe and Theresa Wilson for many insightful discussions. This work was supported in part by National Science Foundation Grants BCS- 0624277 and IIS-0535099 and by DHS Grant N0014-07-1-0152.

References

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

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

Bethard, S., 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.*

Breck, E., Y. Choi, and C. Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of IJCAI.*

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

Choi, Y., E. Breck, and C. Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of EMNLP.*

Choi, F. 2000. Advances in domain independent linear text segmentation. *Proceedings of NAACL.*

Cohen, W. 1995. Fast effective rule induction. In *Proceedings of ICML.*

Freund, Y. and R. Schapire. 1998. Large margin classification using the perceptron algorithm. In *Proceedings of Computational Learing Theory.*

Hasegawa, T., S. Sekine, and R. Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of ACL.*

Hu, M. and B. Liu. 2004. Mining opinion features in customer reviews. In *AAAI.*

Joachims, T. 1998. Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges, A. Smola, editor, *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA.

Kim, S. 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.*

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

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

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

Malioutov, I. and R. Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of ACL/COLING.*

Ng, V. and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *In Proceedings of ACL.*

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

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

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

Rosenfeld, B. and R. Feldman. 2007. Clustering for unsupervised relation identification. In *Proceedings of CIKM.*

Soon, W., H. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4).

Stoyanov, V. and C. Cardie. 2008. Annotating topics of opinions. In *Proceedings of LREC.*

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

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

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

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

Wiebe, J. 2005. Personal communication.

Wilson, T., J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP.*

Wilson, T. 2005. Personal communication.

Yi, J., 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.*