Toward Opinion Summarization: Linking the Sources

Veselin Stoyanov and **Claire Cardie**
Department of Computer Science
Cornell University
Ithaca, NY 14850, USA
{ves,cardie}@cs.cornell.edu

Abstract

We target the problem of linking source mentions that belong to the same entity (source coreference resolution), which is needed for creating opinion summaries. In this paper we describe how source coreference resolution can be transformed into standard noun phrase coreference resolution, apply a state-of-the-art coreference resolution approach to the transformed data, and evaluate on an available corpus of manually annotated opinions.

1 Introduction

Sentiment analysis is concerned with the extraction and representation of attitudes, evaluations, opinions, and sentiment from text. The area of sentiment analysis has been the subject of much recent research interest driven by two primary motivations. First, there is a desire to provide applications that can extract, represent, and allow the exploration of opinions in the commercial, government, and political domains. Second, effective sentiment analysis might be used to enhance and improve existing NLP applications such as information extraction, question answering, summarization, and clustering (e.g. Riloff et al. (2005), Stoyanov et al. (2005)).

Several research efforts (e.g. Riloff and Wiebe (2003), Bethard et al. (2004), Wilson et al. (2004), Yu and Hatzivassiloglou (2003), Wiebe and Riloff (2005)) have shown that sentiment information can be extracted at the sentence, clause, or individual opinion expression level (*fine-grained opinion information*). However, little has been done to develop methods for combining fine-grained opinion information to form a summary representation in which expressions of opinions from the same source/target[1] are grouped together, multiple opinions from a source toward the same target are accumulated into an aggregated opinion, and cumulative statistics are computed for each source/target. A simple opinion summary[2] is shown in Figure 1. Being able to create opinion summaries is important both for stand-alone applications of sentiment analysis as well as for the potential uses of sentiment analysis as part of other NLP applications.

In this work we address the dearth of approaches for summarizing opinion information. In particular, we focus on the problem of *source coreference resolution*, i.e. deciding which source mentions are associated with opinions that belong to the same real-world entity. In the example from Figure 1 performing source coreference resolution amounts to determining that *Stanishev*, *he*, and *he* refer to the same real-world entities. Given the associated opinion expressions and their polarity, this source coreference information is the critical knowledge needed to produce the summary of Figure 1 (although the two target mentions, *Bulgaria* and *our country*, would also need to be identified as coreferent).

Our work is concerned with fine-grained expressions of opinions and assumes that a system can rely on the results of effective opinion and source extractors such as those described in Riloff and Wiebe (2003), Bethard et al. (2004), Wiebe and Riloff (2005) and Choi et al. (2005). Presented with sources of opinions, we approach the problem of source coreference resolution as the closely

[1] We use *source* to denote an opinion holder and *target* to denote the entity toward which the opinion is directed.

[2] For simplicity, the example summary does not contain any source/target statistics or combination of multiple opinions from the same source to the same target.

Proceedings of the Workshop on Sentiment and Subjectivity in Text, pages 9–14,
Sydney, July 2006.

" [Target Delaying of Bulgaria's accession to the EU] would be a *serious mistake*" [Source Bulgarian Prime Minister Sergey Stanishev] said in an interview for the German daily Suddeutsche Zeitung. "[Target Our country] *serves as a model and encourages* countries from the region to follow despite the difficulties", [Source he] added.

[Target Bulgaria] is *criticized* by [Source the EU] because of slow reforms in the judiciary branch, the newspaper notes.

Stanishev was elected prime minister in 2005. Since then, [Source he] has been a *prominent supporter* of [Target his country's accession to the EU].

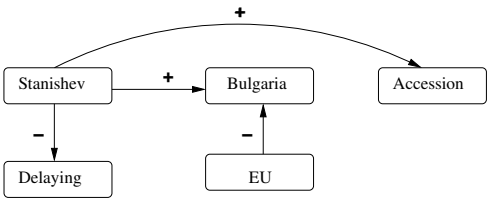

Figure 1: Example of text containing opinions (above) and a summary of the opinions (below). In the text, sources and targets of opinions are marked and opinion expressions are shown in italic. In the summary graph, + stands for positive opinion and - for negative.

related task of noun phrase coreference resolution. However, source coreference resolution differs from traditional noun phrase (NP) coreference resolution in two important aspects discussed in Section 4. Nevertheless, as a first attempt at source coreference resolution, we employ a state-of-the-art machine learning approach to NP coreference resolution developed by Ng and Cardie (2002). Using a corpus of manually annotated opinions, we perform an extensive evaluation and obtain strong initial results for the task of source coreference resolution.

2 Related Work

Sentiment analysis has been a subject of much recent research. Several efforts have attempted to automatically extract opinions, emotions, and sentiment from text. 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 (2002), Dave et al. (2003), Pang and Lee (2004)). In contrast, the problem of fine-grained opinion extraction has concentrated on recognizing opinions at the sentence, clause, or individual opinion expression level. Recent work has shown that systems can be trained to recognize opinions, their polarity, and their strength at a reasonable degree of accuracy (e.g. Dave et al. (2003), Riloff and Wiebe (2003), Bethard et al. (2004), Pang and Lee (2004), Wilson et al. (2004), Yu and Hatzivassiloglou (2003), Wiebe and Riloff (2005)). Additionally, researchers have been able to effectively identify sources of opinions automatically (Bethard et al., 2004; Choi et al., 2005; Kim and Hovy, 2005). Finally, Liu et al. (2005) summarize automatically generated opinions about products and develop interface that allows the summaries to be vizualized.

Our work also draws on previous work in the area of coreference resolution, which is a relatively well studied NLP problem. Coreference resolution is the problem of deciding what noun phrases in the text (i.e. *mentions*) refer to the same real-world entities (i.e. *are coreferent*). Generally, successful approaches have relied machine learning methods trained on a corpus of documents annotated with coreference information (such as the MUC and ACE corpora). Our approach to source coreference resolution is inspired by the state-of-the-art performance of the method of Ng and Cardie (2002).

3 Data set

We begin our discussion by describing the data set that we use for development and evaluation.

As noted previously, we desire methods that work with automatically identified opinions and sources. However, for the purpose of developing and evaluating our approaches we rely on a corpus of manually annotated opinions and sources. More precisely, we rely on the MPQA corpus (Wilson and Wiebe, 2003)[3], which contains 535 manually annotated documents. Full details about the corpus and the process of corpus creation can be found in Wilson and Wiebe (2003); full details of the opinion annotation scheme can be found in Wiebe et al. (2005). For the purposes of the discussion in this paper, the following three points suffice.

First, the corpus is suitable for the domains and genres that we target – all documents have occurred in the world press over an 11-month period, between June 2001 and May 2002. Therefore, the

[3]The MPQA corpus is available at http://nrrc.mitre.org/NRRC/publications.htm.

corpus is suitable for the political and government domains as well as a substantial part of the commercial domain. However, a fair portion of the commercial domain is concerned with opinion extraction from product reviews. Work described in this paper does not target the genre of reviews, which appears to differ significantly from newspaper articles.

Second, all documents are manually annotated with phrase-level opinion information. The annotation scheme of Wiebe et al. (2005) includes phrase level opinions, their sources, as well as other attributes, which are not utilized by our approach. Additionally, the annotations contain information that allows coreference among source mentions to be recovered.

Finally, the MPQA corpus contains no coreference information for general NPs (which are not sources). This might present a problem for traditional coreference resolution approaches, as discussed throughout the paper.

4 Source Coreference Resolution

In this Section we define the problem of source coreference resolution, describe its challenges, and provide an overview of our general approach.

We define *source coreference resolution* as the problem of determining which mentions of opinion sources refer to the same real-world entity. Source coreference resolution differs from traditional supervised NP coreference resolution in two important aspects. First, sources of opinions do not exactly correspond to the automatic extractors' notion of noun phrases (NPs). Second, due mainly to the time-consuming nature of coreference annotation, NP coreference information is incomplete in our data set: NP mentions that are not sources of opinion are not annotated with coreference information (even when they are part of a chain that contains source NPs)[4]. In this paper we address the former problem via a heuristic method for mapping sources to NPs and give statistics for the accuracy of the mapping process. We then apply state-of-the-art coreference resolution methods to the NPs to which sources were mapped (*source noun phrases*). The latter problem of developing methods that can work with incomplete supervisory information is addressed in a subsequent effort (Stoyanov and Cardie, 2006).

[4]This problem is illustrated in the example of Figure 1 The underlined *Stanishev* is coreferent with all of the Stanishev references marked as sources, but, because it is used in an objective sentence rather than as the source of an opinion, the reference would be omitted from the *Stanishev* source coreference chain. Unfortunately, this proper noun might be critical in establishing coreference of the final source reference *he* with the other mentions of the source *Stanishev*.

	Single Match	Multiple Matches	No Match
Total	7811	3461	50
Exact	6242	1303	0

Table 1: Statistics for matching sources to noun phrases.

Our general approach to source coreference resolution consists of the following steps:

1. **Preprocessing:** We preprocess the corpus by running NLP components such as a tokenizer, sentence splitter, POS tagger, parser, and a base NP finder. Subsequently, we augment the set of the base NPs found by the base NP finder with the help of a named entity finder. The preprocessing is done following the NP coreference work by Ng and Cardie (2002). From the preprocessing step, we obtain an augmented set of NPs in the text.

2. **Source to noun phrase mapping:** The problem of mapping (manually or automatically annotated) sources to NPs is not trivial. We map sources to NPs using a set of heuristics.

3. **Coreference resolution:** Finally, we restrict our attention to the source NPs identified in step 2. We extract a feature vector for every pair of source NPs from the preprocessed corpus and perform NP coreference resolution.

The next two sections give the details of Steps 2 and 3, respectively. We follow with the results of an evaluation of our approach in Section 7.

5 Mapping sources to noun phrases

This section describes our method for heuristically mapping sources to NPs. In the context of source coreference resolution we consider a noun phrase to correspond to (or match) a source if the source and the NP cover the exact same span of text. Unfortunately, the annotated sources did not always match exactly a single automatically extracted NP. We discovered the following problems:

1. **Inexact span match.** We discovered that often (in 3777 out of the 11322 source mentions) there is no noun phrase whose span matches exactly the source although there are noun phrases that overlap the source. In most cases this is due to the way spans of sources are marked in the data. For instance, in some cases determiners are not included in the source span (e.g. "*Venezuelan people*" vs. "*the Venezuelan people*"). In other cases, differences are due to mistakes by the NP extractor (e.g. "*Muslims rulers*" was not recognized, while "*Muslims*" and "*rulers*" were recognized). Yet in other cases, manually marked sources do not match the definition of a noun phrase. This case is described in more detail next.

	Measure	Overall rank	Method and parameters	Instance selection	B^3	MUC score	Positive Identification			Actual Pos. Identification		
							Prec.	Recall	F1	Prec.	Recall	F1
400 Training Documents	B^3	1	svm C10 γ0.01	none	**81.8**	71.7	80.2	43.7	56.6	57.5	62.9	60.2
		5	ripper asc L2	soon2	**80.7**	72.2	74.5	45.2	56.3	55.1	62.1	58.4
	MUC Score	1	svm C10 γ0.01	soon1	77.3	**74.2**	67.4	51.7	58.5	37.8	70.9	49.3
		4	ripper acs L1.5	soon2	78.4	**73.6**	68.3	49.0	57.0	40.0	69.9	50.9
	Positive identification	1	svm C10 γ0.05	soon1	72.7	73.9	60.0	57.2	**58.6**	37.8	71.0	49.3
		4	ripper acs L1.5	soon1	78.9	73.6	68.8	48.9	**57.2**	40.0	69.9	50.9
	Actual pos. identification	1	svm C10 γ0.01	none	81.8	71.7	80.2	43.7	56.6	57.5	62.9	**60.2**
		2	ripper asc L4	soon2	73.9	69.9	81.1	40.2	53.9	69.8	52.5	**60.0**
200 Training Documents	B^3	1	ripper acs L4	none	**81.8**	67.8	91.4	32.7	48.2	72.0	52.5	60.6
		9	svm C10 γ0.01	none	**81.4**	70.3	81.6	40.8	54.4	58.4	61.6	59.9
	MUC Score	1	svm C1 γ0.1	soon1	74.8	**73.8**	63.2	55.2	58.9	32.1	74.4	44.9
		5	ripper acs L1	soon1	77.9	**0.732**	71.4	46.5	56.3	37.7	69.7	48.9
	Positive identification	1	svm C1 γ0.1	soon1	74.8	73.8	63.2	55.2	**58.9**	32.1	74.4	44.9
		4	ripper acs L1	soon1	75.3	72.4	69.1	48.0	**56.7**	33.3	72.3	45.6
	Actual pos. identification	1	ripper acs L4	none	81.8	67.8	91.4	32.7	48.2	72.0	52.5	**60.6**
		10	svm C10 γ0.01	none	81.4	70.3	81.6	40.8	54.4	58.4	61.6	**59.9**

Table 2: Performance of the best runs. For SVMs, γ stands for RBF kernel with the shown γ parameter.

2. **Multiple NP match.** For 3461 of the 11322 source mentions more than one NP overlaps the source. In roughly a quarter of these cases the multiple match is due to the presence of nested NPs (introduced by the NP augmentation process introduced in Section 3). In other cases the multiple match is caused by source annotations that spanned multiple NPs or included more than only NPs inside its span. There are three general classes of such sources. First, some of the marked sources are appositives such as "*the country's new president, Eduardo Duhalde*". Second, some sources contain an NP followed by an attached prepositional phrase such as "*Latin American leaders at a summit meeting in Costa Rica*". Third, some sources are conjunctions of NPs such as "*Britain, Canada and Australia*". Treatment of the latter is still a controversial problem in the context of coreference resolution as it is unclear whether conjunctions represent entities that are distinct from the conjuncts. For the purpose of our current work we do not attempt to address conjunctions.

3. **No matching NP.** Finally, for 50 of the 11322 sources there are no overlapping NPs. Half of those (25 to be exact) included marking of the word "*who*" such as in the sentence "*Carmona named new ministers, including two military officers* **who** *rebelled against Chavez*". From the other 25, 19 included markings of non-NPs including question words, qualifiers, and adjectives such as "*many*", "*which*", and "*domestically*". The remaining six are rare NPs such as "*lash*" and "*taskforce*" that are mistakenly not recognized by the NP extractor.

Counts for the different types of matches of sources to NPs are shown in Table 1. We determine the match in the problematic cases using a set of heuristics:

1. If a source matches any NP exactly in span, match that source to the NP; do this even if multiple NPs overlap the source – we are dealing with nested NP's.

2. If no NP matches matches exactly in span then:

   - If a single NP overlaps the source, then map the source to that NP. Most likely we are dealing with differently marked spans.
   - If multiple NPs overlap the source, determine whether the set of overlapping NPs include any non-nested NPs. If all overlapping NPs are nested with each other, select the NP that is closer in span to the source – we are still dealing with differently marked spans, but now we also have nested NPs. If there is more than one set of nested NPs, then most likely the source spans more than a single NP. In this case we select the outermost of the last set of nested NPs before any preposition in the span. We prefer: the outermost NP because longer NPs contain more information; the last NP because it is likely to be the head NP of a phrase (also handles the case of explanation followed by a proper noun); NP's before preposition, because a preposition signals an explanatory prepositional phrase.

3. If no NP overlaps the source, select the last NP before the source. In half of the cases we are dealing with the word *who*, which typically refers to the last preceding NP.

6 Source coreference resolution as coreference resolution

Once we isolate the source NPs, we apply coreference resolution using the standard combination of classification and single-link clustering (e.g. Soon et al. (2001) and Ng and Cardie (2002)).

We compute a vector of 57 features for every pair of source noun phrases from the preprocessed corpus. We use the training set of pairwise instances to train a classifier to predict whether a source NP pair should be classified as positive (the NPs refer to the same entity) or negative (different entities). During testing, we use the trained classifier to predict whether a source NP pair is positive and single-link clustering to group together sources that belong to the same entity.

7 Evaluation

For evaluation we randomly split the MPQA corpus into a training set consisting of 400 documents

and a test set consisting of the remaining 135 documents. We use the same test set for all evaluations, although not all runs were trained on all 400 training documents as discussed below.

The purpose of our evaluation is to create a strong baseline utilizing the best settings for the NP coreference approach. As such, we try the two reportedly best machine learning techniques for pairwise classification – RIPPER (for Repeated Incremental Pruning to Produce Error Reduction) (Cohen, 1995) and support vector machines (SVMs) in the SVM^{light} implementation (Joachims, 1998). Additionally, to exclude possible effects of parameter selection, we try many different parameter settings for the two classifiers. For RIPPER we vary the order of classes and the positive/negative weight ratio. For SVMs we vary C (the margin tradeoff) and the type and parameter of the kernel. In total, we use 24 different settings for RIPPER and 56 for SVM^{light}.

Additionally, Ng and Cardie reported better results when the training data distribution is balanced through instance selection. For instance selection they adopt the method of Soon et al. (2001), which selects for each NP the pairs with the n preceding coreferent instances and all intervening non-coreferent pairs. Following Ng and Cardie (2002), we perform instance selection with $n = 1$ ($soon1$ in the results) and $n = 2$ ($soon2$). With the three different instance selection algorithms ($soon1$, $soon2$, and none), the total number of settings is 72 for RIPPER and 168 for SVMa. However, not all SVM runs completed in the time limit that we set – 200 min, so we selected half of the training set (200 documents) at random and trained all classifiers on that set. We made sure to run to completion on the full training set those SVM settings that produced the best results on the smaller training set.

Table 2 lists the results of the best performing runs. The upper half of the table gives the results for the runs that were trained on 400 documents and the lower half contains the results for the 200-document training set. We evaluated using the two widely used performance measures for coreference resolution – MUC score (Vilain et al., 1995) and B^3 (Bagga and Baldwin, 1998). In addition, we used performance metrics (precision, recall and F1) on the identification of the positive class. We compute the latter in two different ways – either by using the pairwise decisions as the classifiers outputs them or by performing the clustering of the source NPs and then considering a pairwise decision to be positive if the two source NPs belong to the same cluster. The second option (marked *actual* in Table 2) should be more representative of a good clustering, since coreference decisions are important only in the context of the clusters that they create.

Table 2 shows the performance of the best RIPPER and SVM runs for each of the four evaluation metrics. The table also lists the rank for each run among the rest of the runs.

7.1 Discussion

The absolute B^3 and MUC scores for source coreference resolution are comparable to reported state-of-the-art results for NP coreference resolutions. Results should be interpreted cautiously, however, due to the different characteristics of our data. Our documents contained 35.34 source NPs per document on average, with coreference chains consisting of only 2.77 NPs on average. The low average number of NPs per chain may be producing artificially high score for the B^3 and MUC scores as the modest results on positive class identification indicate.

From the relative performance of our runs, we observe the following trends. First, SVMs trained on the full training set outperform RIPPER trained on the same training set as well as the corresponding SVMs trained on the 200-document training set. The RIPPER runs exhibit the opposite behavior – RIPPER outperforms SVMs on the 200-document training set and RIPPER runs trained on the smaller data set exhibit better performance. Overall, the single best performance is observed by RIPPER using the smaller training set.

Another interesting observation is that the B^3 measure correlates well with good "actual" performance on positive class identification. In contrast, good MUC performance is associated with runs that exhibit high recall on the positive class. This confirms some theoretical concerns that MUC score does not reward algorithms that recognize well the absence of links. In addition, the results confirm our conjecture that "actual" precision and recall are more indicative of the true performance of coreference algorithms.

8 Conclusions

As a first step toward opinion summarization we targeted the problem of source coreference resolution. We showed that the problem can be tackled effectively as noun coreference resolution.

One aspect of source coreference resolution that we do not address is the use of unsupervised information. The corpus contains many automatically identified non-source NPs, which can be used to benefit source coreference resolution in two ways. First, a machine learning approach could use the unlabeled data to estimate the overall distributions. Second, some links between sources may be realized through a non-source NPs (see the example of figure 1). As a follow-up to the work described in this paper we developed a method that utilizes the unlabeled NPs in the corpus using a structured rule learner (Stoyanov and Cardie, 2006).

Acknowledgements

The authors would like to thank Vincent Ng and Art Munson for providing coreference resolution code, members of the Cornell NLP group (especially Yejin Choi and Art Munson) for many helpful discussions, and the anonymous reviewers for their insightful comments. This work was supported by the Advanced Research and Development Activity (ARDA), by NSF Grants IIS-0535099 and IIS-0208028, by gifts from Google and the Xerox Foundation, and by an NSF Graduate Research Fellowship to the first author.

References

A. Bagga and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of COLING/ACL*.

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*.

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

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*.

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

S. Kim and E. Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI Workshop on Question Answering in Restricted Domains*.

B. Liu, M. Hu, and J. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of International World Wide Web Conference*.

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

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*.

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

E. Riloff, J. Wiebe, and W. Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings of AAAI*.

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

V. Stoyanov and C. Cardie. 2006. Partially supervised coreference resolution for opinion summarization through structured rule learning. In *Proceedings of EMNLP*.

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

P. Turney. 2002. 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 MUC-6*.

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).

T. Wilson and J. Wiebe. 2003. Annotating opinions in the world press. *4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*.

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

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*.