Automatic Extraction of Opinion Propositions and their Holders

Steven Bethard[†], **Hong Yu**[*], **Ashley Thornton**[†], **Vasileios Hatzivassiloglou**[‡,*] and **Dan Jurafsky**[†]

[†]Center for Spoken Language Research, University of Colorado, Boulder, CO 80309
[‡]Center for Computational Learning Systems, Columbia University, New York, NY 10027
[*]Department of Computer Science, Columbia University, New York, NY 10027
{steven.bethard, ashley.thornton, jurafsky}@colorado.edu
{hongyu, vh}@cs.columbia.edu

Abstract

We identify a new task in the ongoing analysis of opinions: finding propositional opinions, sentential complements which for many verbs contain the actual opinion, rather than full opinion sentences. We propose an extension of semantic parsing techniques, coupled with additional lexical and syntactic features, that can produce labels for propositional opinions as opposed to other syntactic constituents. We describe the annotation of a small corpus of 5,139 sentences with propositional opinion information, and use this corpus to evaluate our methods. We also present results that indicate that the proposed methods can be extended to the related task of identifying opinion holders and associating them with propositional opinions.

Introduction

Separating subjective from objective information is a challenging task that impacts several natural language processing applications. Published news articles often contain factual information along with opinions, either as the outcome of analysis or quoted directly from primary sources. Text materials from many other sources (e.g., the web) also mix facts and opinions. Automatically determining which part of these documents is fact and which is opinion would help in selecting the appropriate type of information given an application and in organizing and presenting that information. For example, an information extraction system would likely prioritize factual parts of a document for analysis, while an advanced question answering or summarization system would need to present opinions separately from facts, organized by source and perspective.

This need for identifying opinions has motivated a number of automated methods for detecting opinions or other subjective text passages (Wiebe, Bruce, & O'Hara 1999; Hatzivassiloglou & Wiebe 2000; Wiebe 2000; Wiebe *et al.* 2002; Riloff, Wiebe, & Wilson 2003; Yu & Hatzivassiloglou 2003) and assigning them to subcategories such as positive and negative opinions (Pang, Lee, & Vaithyanathan 2002; Turney 2002; Yu & Hatzivassiloglou 2003). A variety of machine learning techniques have been employed for this purpose, generally based on lexical cues associated with opinions. However, a common element of current approaches is their focus on either an entire document (Pang, Lee, & Vaithyanathan 2002; Turney 2002) or on full sentences (Wiebe, Bruce, & O'Hara 1999; Hatzivassiloglou & Wiebe 2000; Wiebe 2000; Wiebe *et al.* 2002; Yu & Hatzivassiloglou 2003). In this paper, we examine an alternative approach that seeks to determine opinion status for smaller pieces of text, not by reapplying existing techniques to the clause level but by adopting a more analytic interpretation. In this approach, distinct components of opinion sentences are annotated with specific roles relative to the opinion, such as the opinion holder, the topic of this opinion, and the actual subjective part of the opinion sentence, as opposed to additional factual material; often a sentence that contains subjective clauses expresses an opinion only in the main part or one of the clauses.

We define *opinion* as a sentence, or part of a sentence, that would answer the question "How does X feel about Y?" The opinion needs to be directly stated; this does not include inferences that one could make about how a speaker feels based on word choice. Opinions do not include statements verifiable by scientific data nor predictions about the future.

As an example, consider applying our definition of an opinion to the following two sentences:

(1) I believe in the system.

(2) I believe [you have to use the system to change it].

Both (1) and (2) would be considered opinions under our definition—the first answers the question "How does the author feel about the system?", and the second answers the question "How does the author feel about changing the system?" However, in (1), the scope of the opinion is the whole sentence, while in (2) the opinion of the author is contained within the proposition argument of the verb "believe".

In fact, an opinion localized in the propositional argument of certain verbs as in sentence (2) above is a common case of component opinions. We call such opinions *propositional opinions*. A propositional opinion is an opinion that appears as a semantic proposition, generally functioning as the sentential complement of a predicate. For example, in sentences (3)–(5) below, the underlined portions are propositional opinions, appearing as the complements of the predicates *believe*, *realize*, and *reply*:

(3) I *believe* [you have to use the system to change it].

(4) Still, Vista officials *realize* [they're relatively fortunate].

(5) ["I'd be destroying myself"] *replies* Mr. Korotich.

Not all propositions are opinions. Propositions also appear as complements of verbs like *forget*, *know*, *guess*, *imagine*, and *learn*, and many of these complements are not opinions, as the examples below show:

(6) I don't *know* [anything unusual happening here].

(7) I *understand* [that there are studies by Norwegians that show declining UV-B at the surface].

Our goal in this paper is to automatically extract these propositional opinions. Our interest in this task derives from our interest in automatic question answering, and in particular in answering questions about opinions. Answering an opinion question (like "How does X feel about Y?" or "What do people think about Z?") requires finding which clauses express the exact opinion of the subject. Propositional opinions are an extremely common way to express such third-party opinions. In addition to its key role in opinion question answering, solving the problem of extracting propositional opinions would be an excellent first step toward breaking down opinions into their various components. Finally, we chose propositional opinions because the task was a natural extension from one we had already addressed: extraction of propositions and other semantic/thematic roles from text. Semantically annotated databases like FrameNet (Baker, Fillmore, & Lowe 1998) and PropBank (Kingsbury, Palmer, & Marcus 2002) already mark semantic constituents of sentences like AGENT, THEME, and PROPOSITION, which we expected would help in extracting propositional opinions and opinion-holders.

Our technique for extracting propositional opinions augments an algorithm developed in our earlier work on semantic parsing (Gildea & Jurafsky 2002; Pradhan *et al.* 2003) with new lexical features representing *opinion words*. In the semantic parsing work, sentences were labeled for thematic roles (AGENT, THEME, and PROPOSITION among others) by training statistical classifiers on FrameNet and PropBank. In the current work, we use the actual semantic parsing software described in (Pradhan *et al.* 2003), modifying its role labels so that it performs a binary classification (OPINION-PROPOSITION versus NULL). We use words that are associated with opinions as additional features for this model; these words are automatically learned by bootstrapping from smaller sets of known such words. We examine a classifier that directly assigns opinion status to propositions using these features as well as a two-tiered approach that classifies propositions recognized by the semantic parser. Finally, we present results from a three-way classification where we label sentence constituents as either OPINION-PROPOSITION, OPINION-HOLDER, or NULL.

To be able to train our different classification models, we undertook an annotation effort of 5,139 sentences, marking opinion propositions and opinion holders in them. We discuss our data and its annotation, and then present the opinion word sets we used and the methodology by which they were constructed. Our approaches to the detection of propositions are described in detail, followed by the results we obtained. We conclude with a brief discussion of these results and their likely impact on our continued efforts on extracting and labeling opinion components.

Data

We address the problem of extracting propositional opinions as a supervised statistical classification task, based on hand-labeled training and test sets. In order to label data with propositional opinions, we first established a set of labeling instructions, and then drew upon several resources to build a small corpus of propositional-opinion data.

Labels

In each of the hand-labeling tasks, sentences from a corpus were labeled with one of three labels:

- NON-OPINION
- OPINION-PROPOSITION
- OPINION-SENTENCE

In each of these labels, OPINION indicates an opinion as in our definition above. Thus, the label NON-OPINION means any sentence that could not be used to answer a question of the form "How does X feel about Y?" The remaining two labels, OPINION-PROPOSITION and OPINION-SENTENCE both indicate opinions under our definition, but OPINION-PROPOSITION indicates that the opinion is contained in a propositional verb argument, and OPINION-SENTENCE indicates the opinion is outside of such an argument.

For example, the sentence

(8) I *surmise* [PROPOSITION this is because they are unaware of the shape of humans].

would be labeled NON-OPINION because this sentence does not explain how the speaker feels about the topic; it only makes a prediction about it. By contrast, the sentence

(9) [PROPOSITION It makes the system more flexible] *argues* a Japanese businessman.

would be labeled OPINION-PROPOSITION because the propositional argument in this sentence explains how the businessman feels about "it". Finally, an OPINION-SENTENCE contains an opinion, but that opinion does not fit within the proposition. For example:

(10) It might be *imagined* by those who are not themselves Anglican [PROPOSITION that the habit of going to confession is limited only to markedly High churches] but this is not necessarily the case.

Here, the opinion expressed by the author is not "that the habit of going to confession is limited only to markedly High churches", but that the imaginings of non-Anglicans are not necessarily the case. Thus the opinion is not contained within the proposition argument and so the sentence is labeled OPINION-SENTENCE.

It is worth noting that the labels OPINION-PROPOSITION and OPINION-SENTENCE can occasionally occur in the same sentence. For example:

(11) You may sincerely *believe* yourself [PROPOSITION capable of running a nightclub] and as far as the public relations and administration side goes that's probably true.

Here there are two opinions: the listener's, that they are capable of running a nightclub, and the speaker's, that the listener is probably right. The first of these is contained in the proposition, and the second is not.

FrameNet

FrameNet (Baker, Fillmore, & Lowe 1998) is a corpus of over 100,000 sentences which has been selected form the British National Corpus and hand-annotated for predicates and their arguments. In the FrameNet corpus, predicates are grouped into semantic frames around a *target* verb which have a set of semantic roles. For example the Cognition frame includes verbs like *think*, *believe*, and *know*, and roles like COGNIZER and CONTENT. Each of these roles was mapped onto more abstract thematic roles like AGENT and PROPOSITION via hand-written rules as described in (Gildea & Jurafsky 2002), and later modified by our collaborator Valerie Krugler.

We selected a subset of the FrameNet sentences for hand annotation with our opinion labels. As we are concerned primarily with identifying propositional opinions, we took only the sentences in FrameNet containing a verbal argument labeled PROPOSITION. Each of these sentences was then individually annotated with one or more of the labels above. This produced a dataset of 3,041 sentences, 1,910 labeled as NON-OPINION, 631 labeled OPINION-PROPOSITION, and 573 labeled OPINION-SENTENCE.

PropBank

PropBank (Kingsbury, Palmer, & Marcus 2002) is a million word corpus consisting of the Wall Street Journal portion of the Penn TreeBank that was then annotated for predicates and their arguments. Like FrameNet, PropBank gives semantic/thematic labels to the arguments of each predicate. For an earlier project on semantic parsing, the PropBank labels (ARG0, ARG1, ...) were again mapped into the abstract thematic roles (AGENT, PROPOSITION, etc.) by Valerie Krugler and Karen Kipper.

We again selected only a subset of PropBank for hand annotation with our opinion labels. Using the FrameNet data set, we extracted some verb-specific information. For each verb, we measured the frequency with which that verb occurred with an OPINION (PROPOSITION or SENTENCE) label. These statistics gave an idea of how highly a given verb's use correlates with opinion-type sentences.

We then selected a number of verbs that seemed to correlate highly with OPINION sentences, in order to focus further annotation on sentences more likely to contain opinions. Specifically, we selected the verbs:

accuse	*criticize*	*persuade*	*show*
argue	*demonstrate*	*pledge*	*signal*
believe	*doubt*	*realize*	*suggest*
castigate	*express*	*reckon*	*think*
chastise	*forget*	*reflect*	*understand*
comment	*frame*	*reply*	*volunteer*
confirm	*know*	*scream*	

For each of these verbs, we then labeled all of the PropBank sentences containing these verbs as targets, labeling in the same manner as for the FrameNet sentences. This produced a dataset of 2,098 sentences, 1,203 labeled NON-OPINION, 618 labeled OPINION-PROPOSITION, and 390 labeled OPINION-SENTENCE.

Opinion Holders

In addition to labeling propositional opinions, we also report in this paper our initial experiments in labeling the holder of the opinions. Because our focus is on propositional opinions, we are mainly interested in extracting opinion-holders of each OPINION-PROPOSITION. Example (12) below shows a correctly labeled example:

(12) [OPINION-HOLDER You] can *argue* [OPINION-PROPOSITION these wars are corrective].

To create our training and test sets, we took each OPINION-PROPOSITION that we had labeled in the FrameNet and PropBank corpora, and for each one we hand-labeled the OPINION-HOLDER. For efficiency, we used a semi-automated labeling process, relying on the fact that these PropBank and FrameNet sentences had already been labeled for semantic roles like AGENT. We had observed that the vast majority of OPINION-HOLDERs of propositional opinions were the AGENTs of those sentences (as was the case, for example, in (12) above). We thus automatically labeled each AGENT of an OPINION-PROPOSITION as an OPINION-HOLDER, and hand-checked each, correcting all mistakes. For example, (13) shows a sentence in which the AGENT was not in fact the OPINION-HOLDER, and which had to be hand-corrected to mark "these people" as the OPINION-HOLDER.

(13) Why should [AGENT I] *believe* [OPINION-HOLDER these people] [OPINION-PROPOSITION that one small grey lump which they showed me on a screen is a threat to my life]?

In all, only 10% of the OPINION-HOLDERs in PropBank and FrameNet combined turned out not to be AGENTs and had to be corrected.

Not all opinion holders were explicitly mentioned in the sentences. In 72 sentences (6%) the opinion holder was the "speaker", while in 42 (4%) the opinion holder was unlexicalized. For the purposes of scoring our automatic OPINION-HOLDER labeler, these sentences were counted as if there were no OPINION-HOLDER at all.

Opinion-Oriented Words

Previous work indicated that words that associate with opinions are strong clues for determining phrase and sentence-level subjectivity (Wiebe *et al.* 2002; Riloff, Wiebe, & Wilson 2003; Yu & Hatzivassiloglou 2003). We therefore hypothesized that including such *opinion words* as additional features may enhance the performance of our methods for identifying propositional opinions.

Earlier approaches for obtaining opinion words included manual annotation, as well automatic extension of sets of opinion words by relying on frequency counts and expression patterns. We use as our starting set a collection of opinion words identified by Janyce Wiebe, Ellen Riloff, and col-

leagues using the approaches described above. The collection includes 1,286 *strong opinion words* and 1,687 *weak opinion words*. Examples of strong opinion words include *accuse*, *disapproval*, and *inclination*, while weak opinion words include *abandoned*, *belief*, and *commitment*.

We experimented with using either the strong opinion words in that collection or both the strong and weak opinion words together. Additionally, we explored methods to obtain additional, larger sets of opinion words and assign an *opinion score* to each word.

Our first method relies on differences in the relative frequency of a word in documents that are likely to contain opinions versus documents that contain mostly facts. For this task, we used the TREC 8, 9, and 11 text collections, which consist of more than 1.7 million newswire articles. This corpus includes a large number of Wall Street Journal (WSJ) articles, some of which contain additional headings such as *editorial*, *letter to editor*, *business*, and *news*. We extracted 2,877, 1,695, 2,009 and 3,714 articles in each of these categories, and calculated the ratio of relative frequencies for each word in the editorial plus letter to editor versus the news plus business articles (taken to be representative, respectively, of opinion-heavy and fact-heavy documents).

Our second approach used co-occurrence information, starting from a seed list of 1,336 manually annotated semantically oriented adjectives (Hatzivassiloglou & McKeown 1997), which were considered to be opinion words (Wiebe 2000). We then calculated a modified log-likelihood ratio for all words in our TREC corpus depending on how often each word co-occurred in the corpus in the same sentence with the seed words. Using this procedure, we obtained opinion words from all open classes (adjectives, adverbs, verbs, and nouns).

We also used knowledge in WordNet (Miller *et al.* 1990) to substantially filter the number of words labeled as opinion words by the above methods. We built a supervised Naive Bayes classifier that utilizes as features the hypernyms of each word. For training, we manually annotated a randomly selected set of nouns from the TREC corpus with FACT or OPINION labels, and selected from these 500 fact nouns and 500 opinion nouns. We trained a model using the hypernyms of these nouns as features, to obtain a classifier that predicts a FACT or OPINION label for any given noun.

We evaluated the performance of each of these techniques. We used WordNet part-of-speech information to divide the 1,286 *strong opinion words* into 374 adjectives, 119 adverbs, 951 nouns, and 703 verbs, which we then used as our gold standards. We found that different methods are best for different syntactic classes of opinion words. Our first method was appropriate for verbs while the second method worked better for adverbs and nouns. We applied the WordNet filtering technique to the results of the second method for nouns. There was a trade-off for adjectives—the first method resulted in higher recall while the second method resulted in higher precision. We adopted the first method for adjectives after comparing the average of precision and recall obtained by the two methods in an earlier run, using a subset of the 1,286 strong opinion words manually tagged as adjectives. This first set of adjectives was used only for choosing one of the two methods for extending the set, and the first method was subsequently applied to the full set of 374 adjectives identified with WordNet part-of-speech information, as described above. In that manner, we obtained a total of 19,107/14,713, 305/302, 3,188/22,279 and 2,329/1,663 subjective/objective adjectives, adverbs, nouns and verbs, respectively. Our evaluation demonstrated a precision/recall of 58%/47% for adjectives, 79%/37% for adverbs, 90%/38% for nouns, and 78%/18% for verbs.

Identifying Opinion Propositions

Having identified a large number of opinion-oriented words, we considered two approaches to the opinion identification task. The first directly modifies the semantic parser, restricting the target labels to those relevant to opinion propositions and incorporating the opinion words as additional features, but otherwise uses the same machinery to directly assign labels to sentence constituents. The second approach performs the task in two steps: it first uses a version of the semantic parser to obtain generic semantic constituents (such as PROPOSITION) and then classifies propositions as opinions or not.

One-Tiered Architecture

The one-tiered architecture is a constituent-by-constituent classification scheme. That is, for each constituent in the syntactic parse tree of the sentence, we classify that constituent as either OPINION-PROPOSITION or NULL.

As an example, consider the sentence "He replied that he had wanted to", which has the parse tree in Figure 1. In this situation, we consider each node in the tree, e.g. S1, S, NP, PRP, VP, etc., and assign one of the two labels to this node. For this sentence, the correct classification would be to label the SBAR node as OPINION-PROPOSITION, and the remaining nodes as NULL.

To perform this classification, we use the Support Vector Machine (SVM) (Joachims 1998) paradigm proposed in (Pradhan *et al.* 2003) for semantic parsing, in fact making use of the actual semantic parsing code itself. In that paradigm, semantic roles like AGENT, THEME, PROPOSITION, and LOCATION are labeled by training SVM classifiers. Instead of labeling 20 semantic roles, we simply changed the task to label one: OPINION-PROPOSITION. Our classification task was thus a binary one: OPINION-PROPOSITION versus NULL.

For the semantic parsing task, Pradhan et al. used eight features as input to the SVM classifier—the verb, the cluster of the verb, the subcategorization type of the verb, the syntactic phrase type of the potential argument, the head word of the potential argument, the position (before/after) of the potential argument relative to the verb, the syntactic path in a parse tree between the verb and the potential argument, and the voice (active/passive) of the sentence. A detailed description of each of these features is available in (Gildea & Jurafsky 2002).

Our initial experiments used exactly this feature set. In follow-on experiments, we consider several additional features derived mainly from the opinion-oriented words described in the previous section.

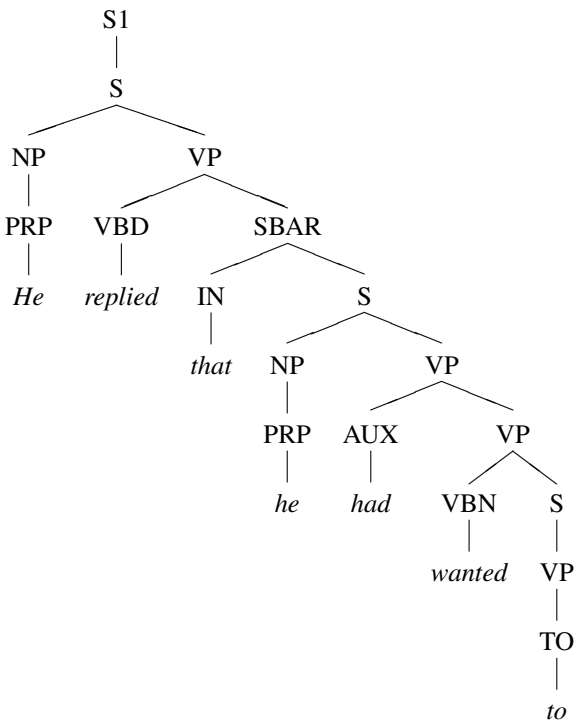

Figure 1: A syntactic parse tree. The SBAR constituent is a propositional opinion.

Counts: This feature counts for each constituent the number of words that occur in a list of opinion-oriented words. We used several alternatives for that list: the strong opinion words identified by Wiebe and colleagues (referred to as "external strong"), both the strong and weak opinion words from that work (referred to as "external strong+weak"), and various subsets of our own automatically constructed list of opinion words obtained by requiring different minimums on each word's opinion score for inclusion in the list.

Score Sum: This feature takes the sum of the opinion scores for each word in the constituent. We again generate several versions of the feature by requiring a different minimum score for inclusion in the total. That is, if we use the feature "Score Sum [Score ≥ 2.0]", we take the sum of all words in the constituent with scores above or equal to 2.0.

ADJP: This is a binary feature indicating whether or not the constituent contains a complex adjective phrase as a child. In exploring our training data, we noticed that adjective phrases with forms like "interested in the idea" seemed to correlate highly with opinions. Simple adjectives, on the other hand, would provide many false positives (e.g., "large" is not likely to be an indicator of opinions). Compare

(14) The accusations were flat and uniform although what is truly remarkable is that the youth of the nation were *believed* [OPINION-PROPOSITION not only to be free of all discipline but also excessively affluent].

and

(15) He felt that shareholder pressure would ensure compliance with the Code but *added* [OPINION-PROPOSITION that if self-regulation does not work a more bureaucratic legislative solution would be inevitable].

which include the underlined complex adjective phrases, with the non-opinion

(16) He *added* [PROPOSITION that there might be a sufficient pool of volunteers to act as a new breed of civil justices].

Dataset	PROPOSITIONAL-OPINION	NULL
Training	912	90,729
Development	178	19,247
Testing	183	19,031

Table 1: Distribution of Constituents as Opinion Propositions or Not.

Using different subsets of these features, we trained several SVM models for labeling propositional opinion constituents. For training and testing data, we selected all the sentences labeled NON-OPINION and all the sentences labeled OPINION-PROPOSITION from both the FrameNet and PropBank datasets. The constituents for propositional arguments in the OPINION-PROPOSITION sentences were labeled as propositional opinions, while all other constituents were labeled NULL.

Some normalization was required to join the two datasets before training our models. First, both FrameNet and PropBank data were stripped of all punctuation as in (Pradhan *et al.* 2003). In addition, propositional arguments in PropBank were slightly altered if they used the complementizer "that". FrameNet labelers were instructed to include "that" in propositional arguments when it occurred as a complementizer, while PropBank labelers were instructed the opposite—"that" was not to be included in the argument. Note that the inclusion of "that" in the argument changes which constituent should receive the propositional-opinion label. Consider the parse tree in Figure 1. The propositional-opinion, as labeled, is shown in the FrameNet style—"that" is included in the proposition—and so the node to receive the label is the SBAR. Under the PropBank labeling style, "that" would not have been included in the proposition, and so the node to receive the label would have been the lower S node. Because our methods learn constituent-by-constituent, it is important to normalize for this sort of labeling so that the data for similar propositional opinion constituents can be shared.

After normalization, both the PropBank and FrameNet data were divided into three randomly selected sets of sentences—70% for training data, 15% for development data, and 15% for testing data. The combined training, development and testing sets were formed by joining the corresponding sets in FrameNet and PropBank. This produced datasets whose sentences were distributed proportionally between FrameNet and PropBank. The distributions of propositional opinion and null constituent labels in each of these datasets are shown in Table 1.

In addition to identifying propositional-opinions, we also considered the task of identifying the holders of these opin-

Dataset	PROPOSITIONAL-OPINION	OPINION-HOLDER	NULL
Training	912	769	89,960
Development	178	149	19,098
Testing	183	162	18,869

Table 2: Distribution of Constituents as Opinion Propositions, Opinion Holders, or Null.

ions. As mentioned above, all OPINION-PROPOSITION sentences were labeled with opinion holders as well. Using the same datasets as above, we trained new models with one additional label: OPINION-HOLDER. The distributions of constituent labels for this three-way classification task are shown in Table 2.

Two-Tiered Architecture

We also explored a two-tiered approach for detecting opinion propositions. The bottom tier was a version of the semantic parser, trained using the Support Vector Machine paradigm proposed in (Pradhan *et al.* 2003) to identify the role of PROPOSITION only (we dropped other semantic roles).

We then built independent classifiers on top of the modified semantic parser to distinguish whether the propositions identified were opinions or not. For this part, we applied our previous machine-learning approach (Yu & Hatzivassiloglou 2003) initially designed for sentence-level opinion and fact classification.

We considered three machine-learning models, all based on a Naive Bayes classifier. The first model trains on approximate labels assigned to the sentence a proposition is part of. These labels are inherited from Wall Street Journal document metadata as described earlier in the section on opinion words; sentences in editorials and letters to the editor are assumed to be opinion sentences, and sentences in news and business articles are assumed to be factual. Predictions are then made for the entire sentence a new proposition is in, and propagated to the individual proposition.

Our second model keeps the training at the sentence level with approximate labels as before, but calculates the predictions only on the text of the proposition which is being classified as opinion or not. Finally, our third model trains directly on propositions using the same kind of approximate, inherited labels, and also predicts on propositions.

All three models use the same set of features which include the words, bigrams, and trigrams in the sentence or proposition, part-of-speech information, and the presence of opinion and positive/negative words; see (Yu & Hatzivassiloglou 2003) for a detailed description of these features. For training the first and second models, we used 20,000 randomly selected sentences from 2,877 editorials and 3,714 news articles from the WSJ. We trained the third model on all 5,147 propositions extracted by the modified semantic parser from these documents. The three models were evaluated on our manually annotated set of propositions in the 5,139 manually annotated sentences from FrameNet and PropBank.

Features	Precision	Recall
No opinion words	50.97%	43.17%
Counts (external, strong)	50.65%	42.62%
Counts (external, strong+weak)	50.00%	43.72%
Counts (Score ≥ 2.0)	52.76%	46.99%
Counts (Score ≥ 2.5)	54.66%	48.09%
Counts (Score ≥ 3.0)	54.27%	48.63%
Score Sum (Score ≥ 0.0)	51.97%	43.17%
Score Sum (Score ≥ 2.0)	52.12%	46.99%
Score Sum (Score ≥ 2.5)	55.35%	48.09%
Score Sum (Score ≥ 3.0)	54.84%	46.45%
ADJP	56.05%	48.09%
ADJP, Score Sum (Score ≥ 2.5)	58.02%	51.37%

Table 3: One-Tiered Approach Results for Opinion Propositions

Results

One-Tiered Architecture

Table 3 shows our results for identifying propositional opinion constituents. The first version of our system used only the features from (Pradhan *et al.* 2003), and no opinion words, and achieved precision of 50.97% and recall of 43.17%.

All of the other systems used at least one of the features presented in our description of the one-tier approach. We found that the counts of subjective words identified in earlier work (our external sets of strong and weak opinion words) were not very good predictors in our task—the systems trained using these features performed nearly identically to the baseline system. The counts of the opinion oriented words identified in the section of this paper on opinion words were better predictors, gaining us, in most cases, several percent (absolute) over the baseline system. We also attempted to take advantage of the scores we produced for these words, and had similar results.

Interestingly, the complex adjective phrase (ADJP) feature provided as much predictive power as the best of our opinion-word based features. Using this feature in combination with our best opinion-oriented word feature, we were able to achieve precision of 58.02% and recall of 51.37%, an 8% (absolute) increase over our baseline for both precision and recall.

Table 4 shows our results for the more difficult, three-way classification into OPINION-PROPOSITION, OPINION-HOLDER, and NULL. Note that the baseline system here performs slightly better than the baseline system in the two-way classification task, while the best system here performs slightly worse than the best two-way system. Still, the results here are are remarkably similar to those we achieved in the easier, two-way classification task which indicates that our system is able to achieve the same performance for propositional opinions and opinion holders as it did for propositional opinions alone.

Two-Tier Architecture

The first step in our two-tier approach is to train a version of the semantic parser using only propositions and target verbs

Train on	Predict on	Measure	Features				
			Words	Bigrams	Trigrams	POS	Orientation
Sentence	Sentence	Recall	33.38%	29.69%	30.09%	30.05%	43.72%
		Precision	67.84%	63.13%	62.50%	65.55%	67.97%
Sentence	Proposition	Recall	37.48%	37.32%	37.79%	36.03%	28.81%
		Precision	53.95%	59.00%	59.83%	55.00%	68.41%
Proposition	Proposition	Recall	42.77%	38.07%	37.84%	35.01%	25.75%
		Precision	59.56%	61.63%	60.43%	58.77%	61.66%

Table 5: Two-tiered Approach Results for Opinion Propositions.

Features	Precision	Recall
No opinion words	53.43%	42.90%
Counts (external, strong)	51.81%	41.45%
Counts (external, strong+weak)	51.04%	42.61%
Counts (Score ≥ 2.0)	54.09%	44.06%
Counts (Score ≥ 2.5)	53.90%	44.06%
Counts (Score ≥ 3.0)	54.93%	45.22%
Score Sum (Score ≥ 0.0)	52.46%	43.19%
Score Sum (Score ≥ 2.0)	54.36%	45.22%
Score Sum (Score ≥ 2.5)	54.74%	45.22%
Score Sum (Score ≥ 3.0)	54.48%	44.06%
ADJP	55.71%	45.22%
ADJP, Score Sum (Score ≥ 2.5)	56.75%	47.54%

Table 4: One-Tiered Approach Results for Opinion Propositions and Opinion Holders

as labels. Our performance in that task was 62% recall and 82% precision, corresponding to an increase of 10% (absolute) in precision over the more general version of the parser with more semantic roles (Pradhan *et al.* 2003).

Table 5 lists the results we obtained by our Naive Bayes classifiers trained over weak, inherited labels from the document level. We generally obtain the highest precision (up to 68%) when we incorporate the opinion / semantic-oriented words in our features. This configuration however usually attains lower recall than just using the words as features, while the bigrams and trigrams offer a slight benefit in most cases. Part-of-speech information did not help either recall or precision. In general, we obtained significantly higher precision values with the two-tier approach as compared to the one-tier approach (68% versus 58%), but at the cost of substantially lower recall (43% versus 51%).

Comparing the three training/prediction models we examined, we note that Model 1 (training and predicting on entire sentences) generally performed better than Models 2 (training on sentences, predicting on propositions) and 3 (training and predicting on propositions). Models 2 and 3 had similar performance. One possible explanation for this difference is that Model 1 uses longer text pieces and thus suffers less from sparse data issues.

Overall, we obtained 43% recall and 68% precision with the best model in the two-tier category, which is lower than our earlier results when we evaluated against manually annotated sentences from the WSJ corpus (Yu & Hatzivassiloglou 2003). The difference between the WSJ (used for training) and BNC corpora (from which our evaluation propositions were drawn) is probably a factor affecting the performance of our two-tier classifiers.

Discussion

We have introduced two new tasks in opinion detection, detecting propositional opinions and the holders of these opinions. While the problem is far from solved, our initial experiments are encouraging. Even these initial experiments have led us to some interesting conclusions. First, the one-tiered and two-tiered approaches offered complementary results, with the one-tiered approach achieving recall and precision of 51%/58% and the two-tiered approach achieving lower recall at a higher precision (43%/68%). Thus, both approaches seem to merit further exploration. Second, our classification was significantly improved by using lists of opinion words which were automatically derived with a variety of statistical methods, and these extended lists proved more useful than smaller, more accurate manually constructed lists. This is a testament to the robustness of those word lists. A new syntactic feature, the presence of complex adjective phrases, also improved the performance of opinion proposition detection. Finally, our results on opinion holder detection show that our approach based on identifying and labeling semantic constituents is promising, and that opinion holders can be identified with accuracy similar to that of opinion propositions.

Acknowledgments

This work was partially supported by a DHS fellowship to the first author, and by ARDA under AQUAINT project MDA908-02-C-0008. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the sponsors. Many thanks to Valerie Krugler and Karen Kipper for mapping roles in the PropBank and FrameNet databases, and to Janyce Wiebe and Ellen Riloff for making available to us their lists of opinion-oriented words. We are also grateful to Ed Hovy, Jim Martin, Kathy McKeown, Rebecca Passonneau, Sameer Pradhan, Wayne Ward, and Janyce Wiebe for many helpful discussions and comments.

References

Baker, C.; Fillmore, C.; and Lowe, J. 1998. The Berkeley FrameNet project. In *Proceedings of the Joint Conference on Computational Linguistics and the 36th Annual Meeting*

of the ACL (COLING-ACL98). Montréal, Canada: Association for Computational Linguistics.

Gildea, D., and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics* **28**(3):245–288.

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, 174–181. Madrid, Spain: Association for Computational Linguistics.

Hatzivassiloglou, V., and Wiebe, J. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the Conference on Computational Linguistics (COLING-2000)*.

Joachims, T. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceeding of the European Conference on Machine Learning*.

Kingsbury, P.; Palmer, M.; and Marcus, M. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*.

Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* **3**(4):235–312.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumps up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*.

Pradhan, S.; Hacioglu, K.; Ward, W.; Martin, J.; and Jurafsky, D. 2003. Semantic role parsing: Adding semantic structure to unstructured text. In *Proceedings of the International Conference on Data Mining (ICDM-2003)*.

Riloff, E.; Wiebe, J.; and Wilson, T. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03)*.

Turney, P. 2002. Thumps up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Wiebe, J.; Wilson, T.; Bruce, R.; Bell, M.; and Martin, M. 2002. Learning subjective language. Technical Report TR-02-100, Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania.

Wiebe, J.; Bruce, R.; and O'Hara, T. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 246–253.

Wiebe, J. 2000. Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*.

Yu, H., and Hatzivassiloglou, V. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*.