Chapter X

Extracting Opinion Propositions and Opinion Holders using Syntactic and Lexical Cues

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
[∞]Department of Linguistics, Stanford university, Stanford CA 94305
Email: {steven.bethard, ashley.thornton}@colorado.edu
Email: {hongyu, vh}@cs.columbia.edu
Email: jurafsky@stanford.edu

Abstract

A new task is identified in the ongoing analysis of opinions: finding propositional opinions, sentential complement clauses of verbs such as "believe" or "claim" that express opinions, and the holders of these opinions. An extension of semantic parsing techniques is proposed that, coupled with additional lexical and syntactic features, can extract these propositional opinions and their opinion holders. A small corpus of 5,139 sentences is annotated with propositional opinion information, and is used for training and evaluation. While our results are still quite preliminary (precisions of 43-51% and recalls of 58-68%), we feel that our focus on opinion clauses, and in general the use of rich syntactic features, helps point to an important new direction in opinion detection.

Keywords: opinions, propositions, semantic parsing, opinion-holders, attribution.

1 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). This chapter examines 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.

In this chapter, an opinion is defined 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 this 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 the 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. In this chapter, such opinions are called 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].

The goal of this chapter is to automatically extract these propositional opinions. An interest in this task derives from 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, this chapter considers propositional opinions because the task was a natural extension from one 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, data which could be expected to help in extracting propositional opinions and opinion-holders.

The technique presented here for extracting propositional opinions augments an algorithm developed in 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 techniques of this chapter, the actual semantic parsing software described in (Pradhan et al., 2003) is used, modifying its role labels so that it performs a binary classification (OPINION-PROPOSITION versus NULL). Words that are associated with opinions are used as additional features for this model; these words are automatically learned by bootstrapping from smaller sets of known such words. A classifier is examined 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, results are presented from a three-way classification where sentence constituents are labeled as either OPINION-PROPOSITION, OPINION-HOLDER, or NULL.

To be able to train different classification models, 5,139 sentences were annotated, marking opinion propositions and opinion-holders in them. This data and its annotation is discussed, and then the opinion word sets used and the methodology by which they were constructed is presented. This chapter's approaches to the detection of propositions are described in detail, followed by the results obtained. A brief discussion of these results and their likely impact on continued efforts on extracting and labeling opinion components concludes the chapter.

2 Data

This chapter addresses 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, a set of labeling instructions was first established, and then several resources were drawn upon to build a small corpus of propositional-opinion data.

2.1 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 the 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 the 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.

2.2 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.

A subset of the FrameNet sentences was selected for hand annotation with opinion labels. As this chapter is concerned primarily with identifying propositional opinions, only the sentences in FrameNet containing a verbal argument labeled PROPOSITION were taken. 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.

2.3 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.

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

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

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

For each of these verbs, all of the PropBank sentences containing these verbs as targets were labeled, 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.

2.4 Opinion-Holders

In addition to labeling propositional opinions, this chapter also reports initial experiments in labeling the holder of the opinions. Because the focus is on propositional opinions, this chapter is 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 training and test sets, each OPINION-PROPOSITION labeled in the FrameNet and PropBank corpora was taken, and for each one an OPINION-HOLDER was hand-labeled. For efficiency, a semi-automated labeling process was used, relying on the fact that these PropBank and FrameNet sentences had already been labeled for semantic roles like AGENT. The vast majority of OPINION-HOLDERs of propositional opinions had been observed to be the AGENTs of those sentences (as was the case, for example, in (12) above). Thus each AGENT of an OPINION-PROPOSITION was automatically labeled as an OPINION-HOLDER, and then hand-checked to correct 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 the automatic OPINION-HOLDER labeler, these sentences were counted as if there were no OPINION-HOLDER at all.

3 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). This chapter therefore hypothesizes that including such opinion words as additional features may enhance the performance of 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. This chapter uses as a starting set a collection of opinion words identified by Janyce Wiebe, Ellen Riloff, and colleagues 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*.

Experiments were performed with using either the strong opinion words in that collection or both the strong and weak opinion words together. Additional methods were explored to obtain additional, larger sets of opinion words and assign an opinion score to each word.

The 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, the TREC 8, 9, and 11 text collections, which consist of more than 1.7 million newswire articles, were used. 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. 2,877, 1,695, 2,009 and 3,714 articles were extracted in each of these categories, and 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) was calculated.

The 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). A modified log-likelihood ratio for all words in the TREC corpus was calculated depending on how often each word co-occurred in the corpus in the same sentence with the seed words. Using this procedure, opinion words were obtained from all open classes (adjectives, adverbs, verbs, and nouns).

Knowledge in WordNet (Miller et al., 1990) was also used to substantially filter the number of words labeled as opinion words by the above methods. A supervised Naïve Bayes classifier was built that utilized as features the hypernyms of each word. For training, a randomly selected set of nouns from the TREC corpus was manually annotated with FACT or OPINION labels, and 500 FACT nouns and 500 OPINION nouns were selected. A model was trained using the hypernyms of these nouns as features, so as to produce a classifier that predicts a FACT or OPINION label for any given noun.

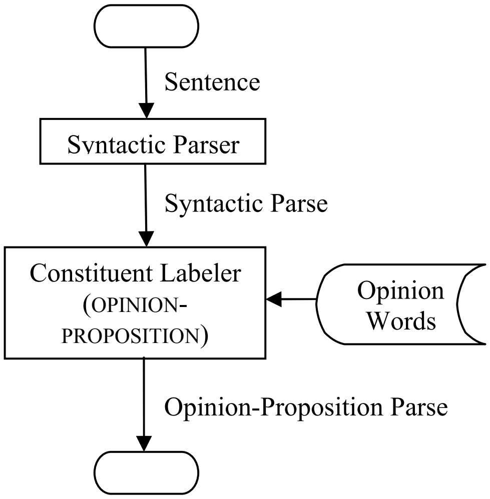
Sentence
Svntactic Parser
Syntactic Parse
Constituent Labeler (OPINION-PROPOSITION)
Opinion Words
Opinion-Proposition Parse

Figure 1: One-tiered architecture

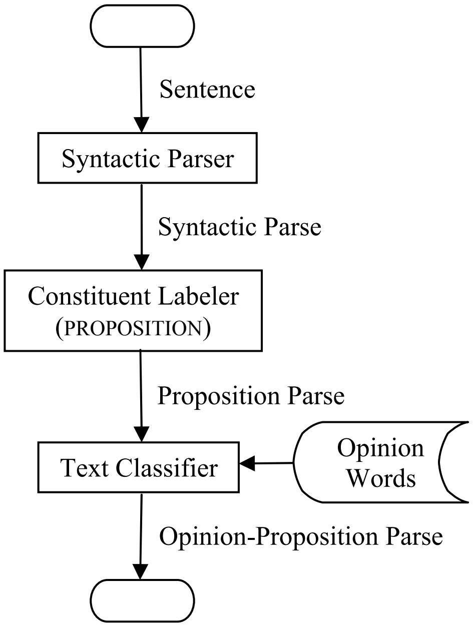

Figure 2: Two-tiered architecture

The performance of each of these techniques was evaluated. WordNet part-of-speech information was used to divide the 1,286 strong opinion words into 374 adjectives, 119 adverbs, 951 nouns, and 703 verbs, which were then used as the gold standards. Different methods proved best for different syntactic classes of opinion words. The first method was appropriate for verbs while the second method worked better for adverbs and nouns. The WordNet filtering technique was applied 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. The first method was adopted 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, 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 were obtained, respectively. The evaluation demonstrated a precision/recall of 58%/47% for adjectives, 79%/37% for adverbs, 90%/38% for nouns, and 78%/18% for verbs.

4 Identifying Opinion Propositions

Having identified a large number of opinion-oriented words, two approaches to the opinion identification task were considered. The first, pictured in Figure 1, 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, pictured in Figure 2, 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.

4.1 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, that constituent is classified as either OPINION-PROPOSITION or NULL.

As an example, consider the sentence "The young Sheikh kept grumbling that the TV was wrong", which has the parse tree in Figure 3. In this situation, each node in the tree, e.g. S1, S, NP, DT, JJ, NNP, VP, etc., is assigned one of the two labels. For this sentence, the correct classification would be to label the SBAR node as OPINION-PROPOSITION, and the remaining nodes as NULL.

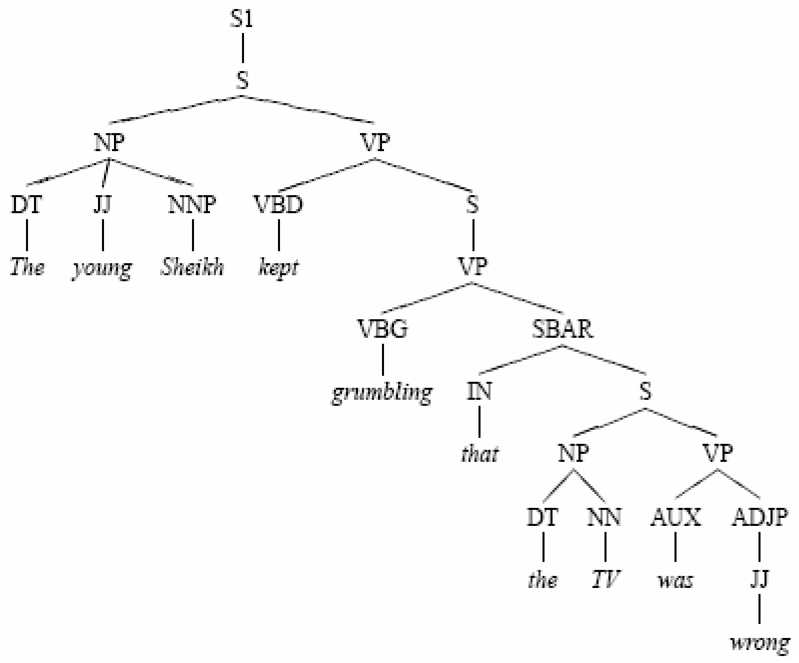

Figure 3: A syntactic parse tree. The SBAR constituent is a propositional opinion.

To perform this classification, the Support Vector Machine (SVM) (Joachims, 1998) paradigm proposed in (Pradhan et al., 2003) for semantic parsing was used, 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, the task was changed to label only one: OPINION-PROPOSITION. The 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).

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

Counts: This feature counts for each constituent the number of words that occur in a list of opinion-oriented words. Several alternatives for that list were considered: 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 the automatically constructed list of opinion words from this chapter, 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. Several versions of the feature were again generated by requiring a different minimum score for inclusion in the total. That is, if the feature "Score Sum [Score ≥ 2.0]" is used, the sum of all words in the constituent with scores above or equal to 2.0 is taken.

ADJP: This is a binary feature indicating whether or not the constituent contains a complex adjective phrase as a child. Explorations of the training data suggested 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 <u>excessively affluent</u>].

and

(15) He felt that shareholder pressure would ensure compliance with the Code but *added* [PROPOSITION that if self-regulation does not work a <u>more bureaucratic</u> 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].

Using different subsets of these features, several SVM models were trained for labeling propositional opinion constituents. For training and testing data, all the sentences labeled NON-OPINION and all the sentences labeled OPINION-PROPOSITION were taken 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 the 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 3. 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 the methods of this chapter 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.

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

Table 1: Distribution of constituents as opinion propositions or null.

In addition to identifying propositional-opinions, the task of identifying the holders of these opinions was also considered. As mentioned above, all OPINION-PROPOSITION sentences were labeled with opinion-holders as well. Using the same datasets as above, new models were trained with one additional label: OPINION-HOLDER. The distributions of constituent labels for this three-way classification task are shown in Table 2.

Dataset	OPINION-PROPOSITION	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.

In addition to treating OPINION-HOLDER as a third label in a single classification task, labeling OPINION-HOLDERs was also approached as a separate task, following the OPINION-PROPOSITION classification task. For this purpose, the sentences that had contained OPINION-PROPOSITIONs were used to train an OPINION-HOLDER vs. NULL constituent classifier.

4.2 Two-Tiered Architecture

A two-tiered approach for detecting opinion propositions was also explored. 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 (other semantic roles were dropped).

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

Three machine-learning models, all based on Naive Bayes learning, were considered. The first model trains on sentences of which 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 labeled to be opinion sentences, and sentences in news and business articles are labeled to be factual. This avoids the need for obtaining individual sentence annotations for training and evaluation, and relies instead on the expectation that documents classified as opinion on the whole (e.g., editorials) will tend to have mostly opinion sentences, and conversely documents placed in the factual category will tend to have mostly factual sentences. Wiebe et al. (2002) report that this expectation is borne out 75% of the time for opinion documents and 56% of the time for factual documents. Predictions are then made for the entire sentence a new proposition is in, and propagated to the individual proposition. The 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, the 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, 20,000 randomly selected sentences from 2,877 editorials and 3,714 news articles from the WSJ were used. The third model was trained on all 5,147 propositions extracted by the modified semantic parser from these documents. The three models were evaluated on the set of opinion propositions manually annotated from FrameNet and PropBank.

5 Results

This section evaluates the one-tiered and two-tiered architectures using the OPINION-PROPOSITION labeled data. For comparison, a baseline system which labels all SBAR constituents as OPINION-PROPOSITIONs, gives a precision of 18.07%, a recall of 50.27%, and an F-score of 26.59%.

5.1 One-Tiered Architecture

Table 3 shows the results for identifying propositional opinion constituents. The first version of the 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 the description of the one-tier approach. The counts of subjective words identified in earlier work (the “external” sets of strong and weak opinion words) were not very good predictors in this task—the systems trained using these features performed nearly identically to the system without them. The counts of the opinion oriented words identified in the section of this chapter on opinion words were better predictors, gaining the system, in most cases, several percent (absolute) in precision and recall. Taking advantage of the scores produced for these words, instead of just their counts, gave similar results.

Interestingly, the complex adjective phrase (ADJP) feature provided as much predictive power as the best of the opinion-word based features. Using this feature in combination with the best opinion-oriented word feature achieved precision of 58.02% and recall of 51.37%, about a 40% (absolute) increase in precision and over the baseline system.

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.

5.1.1 Combined OPINION-PROPOSITION, OPINION-HOLDER Task

Table 4 shows the results for the more difficult, three-way classification into OPINION-PROPOSITION, OPINION-HOLDER, and NULL. Note that the system with no opinion features here performs slightly better than the same 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 remarkably similar to those achieved in the easier, two-way classification task which indicates that the system described here is able to achieve the same performance for propositional opinions and opinion-holders as it did for propositional opinions alone.

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.

5.1.2 Separate OPINION OPINION-PROPOSITION, OPINION-HOLDER Tasks

In the final constituent-labeling experiment, a separate classifier was trained to classify OPINION-HOLDERs after OPINION-PROPOSITIONs had already been classified. In this case, none of the opinion word features were considered because OPINION-HOLDER constituents were expected mainly to contain names or pronouns, not opinion words. For this reason, only one such OPINION-

HOLDER classifier was trained. This classifier was able to identify OPINION-HOLDERs in OPINION-PROPOSITION sentences with 90.85% precision and 89.10% recall. This suggests that once a sentence has been identified as containing an OPINION-PROPOSITION, identifying the OPINION-HOLDER is a much simpler task.

5.2 Two-Tiered Architecture

The first step in the two-tier approach was to train a version of the semantic parser using only propositions and target verbs as labels. 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 obtained by the Naive Bayes classifiers trained over weak, inherited labels from the document level. The highest precision (up to 68%) was generally obtained when the opinion/semantic-oriented words were incorporated as features. This configuration however usually attained lower recall than just using the words as features, while the bigrams and trigrams offered a slight benefit in most cases. Part-of-speech information did not help either recall or precision. In general, significantly higher precision values were obtained 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%).

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.

Comparing the three training/prediction models examined, one notes 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 used longer text pieces and thus suffered less from sparse data issues.

Overall, the best model in the two-tier category obtained 43% recall and 68% precision, a 25% increase in precision and an 18% increase in recall over the baseline system. Still, these results were lower than earlier results that evaluated against manually annotated sentences from the WSJ corpus (Yu & Hatzivassiloglou, 2003). This performance difference is probably due in part to the difference between the WSJ text, which was used for training, and the BNC corpora, from which some of the evaluation propositions were drawn.

6 Error Analysis

In an attempt to find directions for future work, we investigated the errors of our best system, the one tier architecture with results described at the end of Table 3. Overall we found a number of areas where we miss opinion propositions and opinion holders. One key problem was that our current system was based on sentences with punctuation stripped out; it turns out that quotation marks are an important cue to opinion-propositions. We therefore did not detect the opinion-propositions in the following two examples:

(17) [OPINION-PROPOSITION "That must be a comfort,"] *rejoined* [OPINION-HOLDER Ella] as she shut the kitchen door behind her.

(18) [OPINION-PROPOSITION "Liar!"] *snarled* [OPINION-HOLDER her mother]

A related problem was when the opinion was split into two parts, before and after the target; we missed the following split proposition:

(19) [OPINION-PROPOSITION-1 The police] [OPINION-HOLDER he] *concluded* [OPINION-PROPOSITION-2 must possess an unswerving commitment to communication and consultation within which police and the community are equal partners]

We also often missed opinions when the proposition was expression with a noun phrase rather than a full sentential complement, such as the following

(20) [OPINION-HOLDER Mr Chalmers the SNP 's prospective candidate for Glasgow in the European elections] *expressed* [OPINION-PROPOSITION the growing desire within party ranks for an end to the public attacks on the leadership].

Finally, our system is very sensitive to the target opinion verb. Performance on the verb *believe*, for example (P/R=.72/.78) was much higher than average verbs like *argue* (.61/.58), while completely missed detection of opinion clauses (P/R=0/0) for certain verbs (*snarl*, *know*, *chuckle* and *trumpet)*, suggesting obvious directions for improving our system. In addition, our system was quite sensitive to genre; we performed about twice as well on PropBank data as on FrameNet data.

We also examined errors in the opinion holder detection described in section 5.1.1. The major source of errors seemed to be false positives; detecting far too many opinion holders. In many cases we seemed to incorrectly label 2 or even 3 phrases as the opinion holder, caused by the fact that our current architecture makes the opinion-holder decision about each phrase independently, a problem we plan to address. Another such case was when the true opinion holder is the author. We believe these deserve much more attention in the future; the following is such an example:

(21) It does not relieve the need for our market-opening efforts for both goods and services but it does *suggest* [OPINION-PROPOSITION that it is our exports of services and not just borrowing that is financing our imports of goods].

7 Discussion

Two new tasks in opinion detection were introduced: detecting propositional opinions and detecting the holders of these opinions. While these problems are far from solved, the initial experiments of this chapter are encouraging. Even these initial experiments have led 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, 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. In general, syntactic structures based on from the semantic-role-detection literature proved useful. A new syntactic feature, the presence of complex adjective phrases, also improved the performance of opinion proposition detection. Finally, the results on opinion-holder detection show that the 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.

8 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, to Lauren Asia Hall-Lew for help in the error analysis, 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.

9 Bibliography

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)*. Montreal, 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.; andWilson, 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.; 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)*.

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.

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)*.