FINE-GRAINED SUBJECTIVITY AND SENTIMENT ANALYSIS: RECOGNIZING THE INTENSITY, POLARITY, AND ATTITUDES OF PRIVATE STATES

by

Theresa Ann Wilson

Master of Science, University of Pittsburgh, 2001

Submitted to the Graduate Faculty of

School of Arts and Sciences Intelligent Systems Program in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

INTELLIGENT SYSTEMS PROGRAM

This dissertation was presented

by

Theresa Ann Wilson

It was defended on

May 4, 2007

and approved by

Janyce Wiebe, Department of Computer Science

Kevin Ashley, School of Law

Rebecca Hwa, Department of Computer Science

Diane Litman, Department of Computer Science

Dissertation Director: Janyce Wiebe, Department of Computer Science

FINE-GRAINED SUBJECTIVITY AND SENTIMENT ANALYSIS: RECOGNIZING THE INTENSITY, POLARITY, AND ATTITUDES OF PRIVATE STATES

Theresa Ann Wilson, PhD

University of Pittsburgh, 2008

Private states (mental and emotional states) are part of the information that is conveyed in many forms of discourse. News articles often report emotional responses to news stories; editorials, reviews, and weblogs convey opinions and beliefs. This dissertation investigates the manual and automatic identification of linguistic expressions of private states in a corpus of news documents from the world press. A term for the linguistic expression of private states is subjectivity.

The conceptual representation of private states used in this dissertation is that of (Wiebe, Wilson, and Cardie, 2005). As part of this research, annotators are trained to identify expressions of private states and their properties, such as the source and the intensity of the private state. This dissertation then extends the conceptual representation of private states to better model the attitudes and targets of private states. The inter-annotator agreement studies conducted for this dissertation show that the various concepts in the original and extended representation of private states can be reliably annotated.

Exploring the automatic recognition of various types of private states is also a large part of this dissertation. Experiments are conducted that focus on three types of fine-grained subjectivity analysis: recognizing the intensity of clauses and sentences, recognizing the contextual polarity of words and phrases, and recognizing the attribution levels where sentiment and arguing attitudes are expressed. Various supervised machine learning algorithms are used to train automatic systems to perform each of these tasks. These experiments re-

sult in automatic systems for performing fine-grained subjectivity analysis that significantly outperform baseline systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

An important kind of information that is conveyed in many types of written and spoken discourse is the mental or emotional state of the writer or speaker or some other entity referenced in the discourse. News articles, for example, often report individuals' emotional responses to a story in addition to the facts. Editorials, reviews, weblogs, and political speeches convey the opinions, beliefs, or intentions of the writer or speaker. A student engaged in a tutoring session may express his or her understanding or uncertainty. Quirk et al. (1985) give us a general term, **private state**, for referring to these mental and emotional states. In their words, a **private state** is a state that is not open to objective observation or verification: "a person may be observed to *assert that God exists*, but not to *believe that God exists*. Belief is in this sense 'private'." (p. 1181) This dissertation investigates the manual and automatic identification of linguistic expressions of private states in text.

1.1 SUBJECTIVITY ANALYSIS

A term for the linguistic expression of private states, adapted from literary theory (Banfield, 1982), is **subjectivity**. **Subjectivity analysis** is the task of identifying when a private state is being expressed and identifying attributes of the private state. Attributes of private states include who is expressing the private state, the type(s) of attitude being expressed, about whom or what the private state is being expressed, the intensity of the private state, etc. For example, consider the following sentence.

(1.1) The choice of Miers was praised by the Senate's top Democrat, Harry Reid of Nevada.

In this sentence, the phrase "was praised by" indicates that a private state is being expressed. The private state, according to the writer of the sentence, is being expressed by Reid. The attitude being expressed is a positive sentiment, and it is about the choice of Miers, who was nominated to the Supreme Court by President Bush in October 2005.

Subjectivity analysis can also be performed at the sentence level. The goal of sentence-level subjectivity analysis is to determine whether a sentence is **subjective** or **objective**. A sentence is subjective if it contains one or more private state expressions; otherwise, the sentence is objective.

1.2 MOTIVATION FOR FINE-GRAINED SUBJECTIVITY ANALYSIS

Sentence-level subjectivity analysis has proved beneficial for several natural language processing (NLP) tasks. One such task is information extraction (IE), which aims to extract pieces of information that are relevant to an information user. For example, one goal might be to extract from the news information on all terrorist attacks that occurred in the past month. Riloff et. al (2005) showed that identifying subjective sentences and filtering them out can improve information extraction. Another task that has benefited from sentence-level subjectivity analysis is question answering (QA). QA aims to retrieve answers to questions posed in natural language. Although the majority of work in QA has been directed toward developing systems that can answer factoid questions, such as "Who was the first space tourist?," in the past several years, researchers have been pushing the boundaries on the types of questions they hope to be able to answer with QA systems (see Mabury (2004)). One new type of question that researchers in QA are targeting is **opinion questions**. An example of an opinion question is, "How do the Chinese regard the human rights record of the United States?" Automatic subjective sentence filtering has been used by Stoyanov et. al (2005) to improve the ranking of answers to opinion questions.

However, for both IE and QA there is a need for subjectivity analysis below the level of the sentence. Sentences often contain a combination of factual and subjective information. If an IE system is extracting information about terrorist attacks and it discards all subjective

sentences, it will miss the opportunity to extract the information on the attack described in the following subjective sentence.

> (1.2) The Salvadoran government has said that today's attack on the national guard headquarters – in which two children were killed and eight people were injured – shows that the guerrillas "do not want peace and want to escalate violence."

Also, it is not uncommon for more than one private state to be expressed within a single sentence. For example, in the following sentence there are private states expressed for both the President of Iran ("accused the United States of warmongering") and the United States ("American support").

> (1.3) President Mohammad Khatami of Iran, whose attempt at reforms have gotten American support, accused the United States of warmongering.

For a QA system to answer questions about people's opinions, it will need to be able to pinpoint where in a sentence an opinion is being expressed, who is expressing it, and what the opinion is directed toward.

To answer some opinion questions, QA systems will also need to recognize characteristics of the expressed subjectivity. For example, consider the following three opinion questions.

> (Q1) How do Republicans feel about the nomination of Harriet Miers to the Supreme Court?
>
> (Q2) Who has the strongest objections to the nomination of Harriet Miers?
>
> (Q3) What are the opinions about why Harriet Miers decided to withdraw her nomination?

To answer these questions, a QA system will need to be able to recognize different types of attitudes: Q1 and Q2 are asking for sentiments, while Q3 is looking for beliefs or arguments. To answer Q2, the system will need to be able to distinguish between negative sentiments of differing intensities.

In addition to fine-grained subjectivity analysis being needed to extend the capabilities of existing NLP applications, the past few years have seen the development of new applications that also require fine-grained subjectivity analysis. The most prominent of these is mining and summarizing opinions from product reviews (Morinaga et al., 2002; Nasukawa and Yi, 2003; Hu and Liu, 2004). The goal of product review mining is to identify from on-line

reviews what products and product features people judge favorably or unfavorably. This information is then summarized and presented to the user in an informative way. Although review mining is primarily concerned with sentiment, a type of subjectivity, the problem has many of the same characteristics as opinion QA. For example, within a single sentence, there may be both positive and negative sentiments expressed, either toward different aspects of a single product or in comparing one product to another. Also, like other private states, sentiments may vary in intensity. For example, a reviewer may like one product fairly well, but still find another product superior. As with QA, mining and summarizing product reviews requires a system to be able to pinpoint where in a sentence an opinion is being expressed, what type of opinion is being expressed, and what the opinion is directed toward.

In short, there is a need for fine-grained subjectivity analysis to support new NLP applications, such as review mining, and to continue to grow the capabilities of applications such as information extraction and question answering systems.

1.3 RESEARCH OVERVIEW

I have two high-level goals with the research in this dissertation. The first is to contribute to the understanding of how different types of private states are expressed in text, through corpus annotation and analysis. Included in this goal is the extension and development of natural language resources annotated with information about private states. The second high-level goal is to develop automatic systems for performing fine-grained subjectivity analysis, using knowledge gleaned from corpus analysis and the literature.

1.3.1 Recognizing the Intensity, Polarity, and Attitudes of Private States

In this dissertation, I focus on three types of fine-grained subjectivity analysis: recognizing private states of differing intensities, private states of differing polarities, and private states of differing attitudes.

Intensity refers to the strength of the private state that is being expressed, in other words, how strong is an emotion or a conviction of belief. As language users, we intuitively perceive distinctions in the intensity levels of different private states. For example, *outraged* and *extremely annoyed* are more intensely negative than *irritated*. Recognizing intensity includes not only identifying private states of different intensity, but also detecting the absence of private states. Thus, recognizing intensity subsumes the task of distinguishing between subjective and objective language. In this dissertation, I investigate whether human annotators can reliably annotate expressions of private states and their intensities, how private states of differing intensities are expressed, and how sentences and clauses of differing intensity levels may be automatically recognized.

The term **polarity** has a number of different uses, but in this dissertation it is used primarily to refer to the positive or negative sentiment being expressed by a word. However, there is an important distinction between the **prior polarity** of a word and its **contextual polarity**. The **prior polarity** of a word refers to whether a word typically evokes something positive or something negative when taken out of context. For example, the word *beautiful* has a positive prior polarity, and the word *horrid* has a negative prior polarity. The **contextual polarity** of a word is the polarity of the expression in which the word appears, considering the context of the sentence and the discourse. Although words often do have the same prior and contextual polarity, many times the word's prior and contextual polarities differ. Words with a positive prior polarity may have a negative contextual polarity, or vice versa. For example, in sentence 1.4 the negative word "horrid" has a positive contextual polarity. Also, quite often words that are positive or negative out of context are **neutral** in context, meaning that they are not even being used to express a sentiment. This is the case with the word "condemned" in sentence 1.5.

(1.4) Cheers to Timothy Whitfield for the wonderfully horrid (**positive**) visuals.

(1.5) Gavin Elementary School was condemned (**neutral**) in April 2004.

In this dissertation, I investigate how private states of differing polarities are expressed in context, and how this information may be used to automatically recognize the contextual polarity of words and phrases.

Attitudes and their targets are two of the functional components of private states: A private state may be described as the state of an *experiencer*, holding an *attitude*, optionally toward a *target* (Wiebe, 1990; Wiebe, 1994). There are many different kinds of attitudes, for example, sentiments, speculations, beliefs, evaluations, and uncertainty. In this dissertation, I present a conceptual representation of attitude types and their targets, which is an extension to the conceptual representation of private states presented in (Wiebe, 2002; Wiebe, Wilson, and Cardie, 2005). I then investigate whether human annotators can reliably annotate attitudes and targets, and how sentiments and arguing attitudes may be automatically recognized.

1.3.2 General Hypotheses

As I work toward the goals and tasks described above, I explore the following three general hypotheses:

Hypothesis 1: Annotators can be trained to reliably annotate expressions of private states and their attributes.

Hypothesis 2: Automatic systems can be developed for performing fine-grained subjectivity analysis that perform better than baseline systems.

Hypothesis 3: Automatic, fine-grained subjectivity analysis requires a wide variety of features, including both lexical and syntactic clues of subjective language.

I investigate the above hypotheses in different chapters throughout this dissertation. Additionally, in some chapters I investigate more specific hypotheses related to particular studies in fine-grained subjectivity analysis.

1.3.3 Methodology

To understand how different types of private states are expressed in text, I take a corpus linguistics approach. In general, a corpus linguistics approach involves: 1) developing a conceptual representation for the linguistic phenomena of interest, 2) developing coding schemas and manual annotation instructions for the conceptual representation, 3) training

annotators and conducting inter-annotator agreement studies, 4) producing the annotated corpus, and 5) analyzing the corpus to gain insight into how the linguistic phenomena of interest are expressed in context.

Inter-annotator agreement studies are used in this dissertation to test the first general hypothesis. If private states and their attributes can be reliably annotated, trained annotators will be able to achieve acceptable levels of agreement in an annotation study, as measured by standard metrics, such as Cohen's Kappa (κ) (Cohen, 1960) and Krippendorff's Alpha (α) (Krippendorff, 2004). For interpreting κ and α agreement, I use Krippendorff's (1980; 2004) scale, which is the standard that has been adopted by the NLP community. Krippendorff suggests that a κ or α value of 0.80 allows for firm conclusions to be made, and a value of at least 0.67 is sufficient for drawing tentative conclusions. A κ or α of 1 indicates perfect agreement. Thus, for the annotation studies in this dissertation, I consider a κ or α value of 0.67 or higher as evidence supporting the first general hypothesis, with higher values providing stronger evidence.

To develop automatic systems for performing fine-grained subjectivity analysis, I follow a supervised machine learning approach with a focus on feature engineering. Specifically, I use insights into the problem gained from the literature and from corpus analysis to develop linguistically motivated features, for example, features that represent syntactic dependencies that are correlated with particular types of subjectivity. I then use these features in existing machine learning programs to develop the automatic systems.

To test the second general hypothesis, I use an experimental paradigm that involves dividing the corpus into training and testing sets and performing cross-validation experiments using the systems I develop for fine-grained subjectivity analysis. I measure the performance of each system using standard metrics, including accuracy, F-measure, recall, and precision. To evaluate whether a given system performs better than the baseline for a particular task, I use the statistical t-test to test the differences between the average results over the test folds for the automatic system and the average results for the baseline system. I consider a t-test with a p-value < 0.05 to be evidence that a given automatic system performs better than the baseline.

I test the last general hypothesis by conducting ablation experiments and evaluating

the results of automatic systems trained using different sets of features. If it is true that a wide variety of features is required for fine-grained subjectivity analysis, automatic systems that use the widest variety of features will give the best performance. When evaluating features, I consider their performance in systems trained using several different machine learning algorithms. When features improve system performance irrespective of the learning algorithm, this provides strong evidence of the utility of the features.

Taken together, the methodologies and approaches described above form stages in a cycle that is often found within an on-going line of research in natural language processing. In the first stage, a representation is developed for the concept of interest and a corpus is annotated. In the second stage, the corpus is analyzed to gain insight into the concept, and systems are developed to perform automatic recognition or analysis of the concept. At the end of the cycle, the results of at any stage may be analyzed and used to inform the next cycle in the line of research, for example, by leading to refinements or extensions of the original conceptual representation.

1.4 CONTRIBUTIONS OF THIS WORK

The research in this dissertation contributes to an on-going line of research in subjectivity analysis, which began with the work of Wiebe (1990; 1994). In that work, Wiebe connected the concepts of private state and linguistic subjectivity and developed the basic conceptual representation for private states. The line of research continues in (Wiebe, Bruce, and O'Hara, 1999; Bruce and Wiebe, 1999), where Wiebe and colleagues developed a corpus with sentence-level subjectivity annotations, and used this corpus to develop a system for automatically recognizing subjective sentences. In (Wiebe, 2002; Cardie et al., 2003), the conceptual representation for private states was extended to include attributions, to distinguish between different types of private state expressions, and to include additional attributes of private states. My dissertation research begins at this point.

A key contribution of the research in this dissertation is the production of the Multi-perspective Question Answering (MPQA) Opinion Corpus. Beginning with the conceptual

representation for private states in (Wiebe, 2002) and a coding schema, I developed manual annotation instructions for performing the annotations and compiled additional training material. I trained the annotators who annotated the corpus, and, with the first inter-annotator agreement study in this dissertation, I validate that key aspects of the conceptual representation for private states can be annotated reliably[1]. This agreement study is the first successful study to be reported for fine-grained private state judgments, and the MPQA Corpus is the only corpus with detailed, expression-level private state annotations yet to be produced. The MPQA Corpus version 1.0 was release to the research community in the fall of 2003.

As part of this dissertation, I also extend the original conceptual representation (Wiebe, 2002) to better model the attitudes and targets of private states. My extensions to the representation are presented in Chapter 7. I developed annotation instructions and trained an annotator to produce the new annotations. With the second inter-annotator agreement study in this dissertation, I validate that given the original private state annotations attitudes and their targets can be reliably annotated. A version of the MPQA Corpus with the new attitude and target annotations on a large subset of the documents also will be released to the research community.

Using the annotations in the MPQA Corpus, I conduct experiments in the automatic recognition of subjectivity. The experiments focus on recognizing different types of fine-grained subjectivity, and in doing so, push into new and challenging areas of subjectivity research. My experiments in intensity classification are the first to automatically classify the intensity of private states, and the first to perform subjectivity analysis at the clause level for all the sentences in a corpus. These experiments and their results are reported in (Wilson, Wiebe, and Hwa, 2004; Wilson, Wiebe, and Hwa, 2006).

With my experiments in recognizing contextual polarity, I use a two-step procedure to investigate what features are useful not only for recognizing whether a word in context is positive or negative, but also whether a word is neutral as opposed to positive or negative.

[1] Agreement for different aspects of the original conceptual representation are presented as parts of multiple papers. Agreement for core concepts in the representation are first presented in (Wilson and Wiebe, 2003) and later included in (Wiebe, Wilson, and Cardie, 2005). Agreement for another attribute, the intensity of private states, is presented in (Wilson, Wiebe, and Hwa, 2006).

Many of the features that I use in these experiments are new, motivated by ideas borrowed from the linguistics literature and by analysis of the corpus. These experiments are also the first to explore how neutral instances affect the performance of features for distinguishing between positive and negative contextual polarity. An early version of the contextual polarity experiments and their results are reported in (Wilson, Wiebe, and Hoffmann, 2005).

The experiments in attitude-type recognition focus on classifying two types of attitude, sentiment and arguing. Although there has been previous research in recognizing sentiment, these are the first experiments to automatically classify arguing attitudes. They are also the first experiments to classify the attitude of all **attribution levels** in a sentence[2].

1.5 OUTLINE

The MPQA Corpus is used in the annotation studies and experiments throughout this dissertation. However, different versions of the corpus are created or used in each chapter. Thus, in Chapter 2 I give a brief overview of the MPQA Corpus, the changes in each version of the corpus, and which versions are used in the various chapters of this dissertation.

In Chapter 3, I describe the conceptual representation of private states that forms the foundation of this work. I also present in this chapter the inter-annotator agreement study that I conducted to verify that the core aspects of the representation can be reliably annotated.

Chapter 4 gives an overview of the machine learning algorithms that are used in the experiments in later chapters. Chapter 5 presents my experiments in recognizing the intensity of sentences and clauses, and my experiments in recognizing contextual polarity are presented in Chapter 6.

In Chapter 7, I present my conceptual representation for the attitudes and targets of private states. This chapter also presents the inter-annotator agreement study that I conducted to verify that the attitudes and targets of private states can be reliably annotated.

[2]In general, there is an attribution level for the writer of a sentence, as well as attribution levels for every direct or indirect quotation, and every entity experiencing a private state within the sentence. Attribution levels often correspond to clauses, but not all clauses correspond to attribution levels.

My experiments in recognizing attitude types are presented in Chapter 8.

In Chapter 9, I briefly explore what potential fine-grained subjectivity analysis may have for helping with question answering in the OpQA Corpus.

Rather than trying to cover in a single chapter all the work related to the research in this dissertation, I instead chose to devote a section in Chapters 3, 5–8 to the work most relevant to the research presented in each chapter. Then, in Chapter 10 I give a review of research in subjectivity and sentiment analysis.

In Chapter 11 I conclude and discuss directions for further research.

2.0 CORPUS

The Multi-Perspective Question Answering (MPQA) Opinion Corpus is a corpus of news documents from the world press collected as part of the summer 2002 NRRC Workshop on Multi-Perspective Question Answering (Wiebe et al., 2003). The corpus contains 535 documents from 187 different news sources, dating from June 2001 to May 2002. About two thirds of the documents in the MPQA Corpus were selected to be on one of ten specific topics. These topics are listed in Table 2.1.

The MPQA Corpus is the platform for the annotations and experiments in this dissertation. Different versions of the annotated corpus are used in different chapters of this dissertation. Below I describe these versions and give the chapters in which they are used.

MPQA Corpus version 1.0: For this initial version of the corpus, GATE 1.2 (Cunningham et al., 2002) was used to provide word tokenization, part-of-speech tagging, and sentence splitting. The result of this initial processing was a corpus of 10,656 sentences and approximately 265,000 words. Chapter 3 describes the private state annotations in this first version of the corpus, which my research group released to the research community in the fall of 2003. I use version 1.0 in the experiments in Chapter 5 on recognizing the intensity of private states.

MPQA Corpus version 1.2: This second version of the corpus contains corrected sentence splits and the contextual polarity annotations described in Chapter 6.[1] The sentence splitter in GATE 1.2 produced some very blatant errors. These incorrect sentence splits and any private state annotations that they affected were hand-corrected by me

[1]There was also a change in terminology from version 1.0 to version 1.2. However the two terminologies are equivalent, and the representations are homomorphic. Throughout this dissertation, I use the newer terminology.

Table 2.1: Topics in the MPQA Corpus

Topic	Description
argentina	Economic collapse in Argentina
axisofevil	U.S. President's State of the Union Address
guantanamo	Detention of prisoners in Guantanamo Bay
humanrights	U.S. State Department Human Rights Report
kyoto	Kyoto Protocol ratification
settlements	Israeli settlements in Gaza and the West Bank
space	Space missions of various countries
taiwan	Relationship between Taiwan and China
venezuela	Presidential coup in Venezuela
zimbabwe	Presidential election in Zimbabwe

and another annotator. As a result, this version of the corpus has 11,114 sentences. This version also contains a few additional corrections to the original private state annotations. I use the annotations in this version of the corpus in the experiments on recognizing contextual polarity in Chapter 6. My research group released version 1.2 to the public in December 2005.

MPQA Corpus version 2.0: This version of the corpus contains the attitude and target annotations described in Chapter 7. To date, 344 documents (5,957 sentences) have been annotated with attitudes and their targets. I use a large subset of documents with these annotations (303 documents) in the attitude recognition experiments in Chapter 8. Version 2.0 of the corpus is scheduled for release in the spring of 2008.

3.0 REPRESENTING PRIVATE STATES

The conceptual representation for private states (Wiebe, 2002; Wiebe, Wilson, and Cardie, 2005) forms the foundation for the work in this dissertation. The intensity of private states, the focus of Chapter 5, is part of this representation, and the annotations based on this representation are used for training and evaluation in experiments throughout this dissertation. It is this representation that I extend in Chapter 7 with a conceptual representation for attitude types and their targets.

The first part of this chapter describes the conceptual representation for private states. The second part of this chapter then focuses on realizing the conceptualization. I briefly describe how the annotation scheme was implemented and the process for training annotators. Following that, I present the inter-annotator agreement studies that I conducted to verify that private states can be reliably annotated. The chapter ends with a discussion of related work.

3.1 CONCEPTUAL REPRESENTATION

In (Wiebe, 1990; Wiebe, 1994), Wiebe presents a basic representation for private states based on their functional components. Specifically, a private state is described as the state of an *experiencer*, holding an *attitude*, optionally toward a *target*. This basic representation is adapted and expanded in (Wiebe, 2002), and in (Wiebe, Wilson, and Cardie, 2005), the terminology for some of the concepts is updated. The end result is a frame-style conceptual representation of private states and attributions.

The conceptual representation has a total of four representational frame: two types of

private state frames, a frame for objective speech events, and a frame for agents. I describe these frames and their attributes below.

3.1.1 Private State Frames

There are several main ways that private states are expressed in language. Private states may be explicitly mentioned, as with the phrase "have doubts" in sentence 3.1.

(3.1) Democrats also have doubts about Miers' suitability for the high court.

Private states may also be expressed in speech events. In this research, **speech event** is used to refer to any speaking or writing event. Examples of speech events in which private states are being expressed are "was criticized" in sentence 3.2 and "said" in sentence 3.3.

(3.2) Miers' nomination was criticized from people all over the political spectrum.

(3.3) "She [Miers] will be a breath of fresh air for the Supreme Court," LaBoon said.

In 3.2, the word "criticized" is used to convey that a negative evaluation was expressed by many people, without giving their exact words; it implies a mixture of private state and speech. With "said" in 3.3, it is the quoted speech that conveys the private state of the speaker. Specifically, LaBoon uses the phrase "a breath of fresh air" to express his private state. This phrase is an example of an **expressive subjective element** (Banfield, 1982). Expressive subjective elements indirectly express private states, through the way something is described or through a particular wording. Sentence 3.4 contains another example of an expressive subjective element, the phrase "missed opportunity of historic proportions."

(3.4) This [the nomination of Miers] is a missed opportunity of historic proportions.

Private states may also be expressed through **private state actions** (Wiebe, 1994). Booing, sighing, stomping away in anger, laughing, and frowning are all examples of private state actions, as is "applaud" in sentence 3.5.

(3.5) As the long line of would-be voters marched in, those near the front of the queue began to spontaneously applaud those who were far behind them.

Figure 3.1: Direct subjective frame

- **text anchor:** a pointer to the span of text that represents the explicit mention of a private state, speech event expressing a private state, or private state action.
- **source:** the person or entity that expresses or experiences the private state, possibly the writer.
- **target:** the target or topic of the private state, i.e., what the speech event or private state is about.
- **properties:**
  - **intensity:** the intensity of the private state; values *low*, *medium*, *high*, or *extreme*.
  - **expression intensity:** the contribution of the speech event or private state expression itself to the overall intensity of the private state; values *neutral*, *low*, *medium*, *high*, or *extreme*.
  - **implicit:** true, if the frame represents an an *implicit* speech event.
  - **insubstantial:** true, if the private state or speech event is not substantial in the discourse. For example, a private state in the context of a conditional often has the value *true*.
  - **attitude type:** the polarity of the private state; values *positive*, *negative*, *other*, or *none*.

In the conceptual representation, the different ways of expressing private states are represented using two types of private state frames. **Direct subjective frames** are used to represent explicit mentions of private states, speech events expressing private states, and private state actions. **Expressive subjective element frames** are used to represent expressive subjective elements. The two private state frames and their attributes are given in Figures 3.1 and 3.2. Figure 3.3 gives the private state frames for sentences 3.1–3.3. The two types of private state frames share many of the same attributes: *text anchor*, *source*, *intensity*, and *attitude type*. The additional attributes in the direct subjective frame reflect its greater complexity.

As its name indicates, the *text anchor* attribute points to the span of text where the private state frame annotation is anchored. For direct subjective annotations, the text anchor is the phrase for the explicit mention of the private state, the speech event, or the private state action. For expressive subjective elements, the text anchor is the subjective or expressive phrase. However, for speech events that are implicit there is no speech event

Figure 3.2: Expressive subjective element frame

- **text anchor:** a pointer to the span of text that denotes the subjective or expressive phrase.
- **source:** the person or entity that is expressing the private state, possibly the writer.
- **properties:**
  - **intensity:** the intensity of the private state; values *low*, *medium*, *high*, or *extreme*.
  - **attitude type:** the polarity of the private state; values *positive*, *negative*, *other*, or *none*.

phrase on which to anchor the annotation frame. **Implicit speech events** are speech events for which there is not a discourse parenthetical, such as, "she said." For example, every sentence in a document is an implicit speech event for the writer of the document. Similarly, direct quotations are not always accompanied by discourse parentheticals, such as in the second sentence in the following passage:

> (3.6) "It could well be that she is in the tradition of Clarence Thomas or Antonin Scalia, as the president has promised," said Jan LaRue, chief counsel of Concerned Women for America. "The problem is that those of us who were looking for some tangible evidence of that have none, and we can't come out of the box supporting her."

When the speech event for a direct subject annotation is implicit, such as for the writer's subjective speech event above in sentence 3.4, the text anchor points to the sentence or quoted string with the text of the speech event, and the *implicit* attribute is used to mark the annotation.

The *source* attribute is used to mark the experiencer of the private state or the speaker or writer of the speech event. Obviously, the writer of an article is a source, because he or she wrote the sentences that constitute the article. However, the writer may also write about other people's private states and speech events, leading to multiple sources in a single sentence. For example, in sentence 3.1 above, there are two sources, the writer, and Democrats (the experiencer of the private state "have doubts"). There are also two sources in sentences 3.2 and 3.3, respectively. A key aspect of sources is that they are nested to capture the levels of attribution. In 3.1, the Democrats do not directly state that they have doubts. Rather it

Figure 3.3: Private state annotations for example sentences 3.1–3.3

Sentence 3.1 Democrats also **have doubts** about Miers' suitability for the high court.

Direct Subjective
text anchor: *have doubts*
source: <writer, Democrats>
target: Miers
intensity: medium
expression intensity: medium
attitude type: negative

Sentence 3.2 Miers' nomination **was criticized** from people all over the political spectrum.

Direct Subjective
text anchor: *was criticized*
source: <writer, people>
target: Miers
intensity: medium
expression intensity: medium
attitude type: negative

Sentence 3.3 "She [Miers] will be **a breath of fresh air** for the Supreme Court," LaBoon **said**.

Direct Subjective
text anchor: *said*
source: <writer, LaBoon>
target: Miers
intensity: medium
expression intensity: neutral
attitude type: positive

Expressive Subjective Element
text anchor: *a breath of fresh air*
source: <writer, LaBoon>
intensity: medium
attitude type: positive

is according to the writer that the Democrats have doubts about Miers' suitability for the Supreme Court. The full source of the private state expressed by "have doubts" is thus the **nested source**: ⟨*writer, Democrats*⟩. The nested source is composed of IDs associated with each source. These IDs are described below in the section on **agent frames** (Section 3.1.3).

The *intensity* attribute is used to mark the overall intensity of the private state that is represented by the direct subjective or expressive subjective element frame. Its values are *low*, *medium*, *high*, and *extreme*. For direct subjective frames, there is an additional intensity rating. The *expression intensity* attribute is used to mark the contribution to the overall intensity made just by the private state or speech event phrase. The values of this attribute are *neutral*, *low*, *medium*, *high*, and *extreme*. For example, *say* is often neutral, even if what is uttered is not neutral. The word *excoriate*, on the other hand, by itself implies a very strong private state.

To help clarify the differences among the various intensity attributes, consider the annotations for sentences 3.2 and 3.3, which are given in Figure 3.3. In sentence 3.2, there is a direct subjective frame for "was criticized." The intensity of "was criticized" is marked as medium, as is the expression intensity. Sentence 3.3 contains both an expressive subjective element frame and a direct subjective frame. The intensity of "a breath of fresh air" is marked as medium. The intensity for "said" is also marked as medium. This is because for direct subjective frames, the speech event or private state phrase and everything inside the scope of the speech event or private state attributed to the same nested source is considered when judging the overall intensity. The expression intensity for "said" is neutral because the word "said" itself does not contribute to the intensity of the private state.

The *attitude type* attribute is for representing the polarity of the private state.

The *target* attribute is for marking the target or topic of the private state, for example, what the speech event or private state is about.

The *insubstantial* attribute is used to mark direct subjective annotations that are not substantial in the discourse. A private state or speech event may be insubstantial either because it is not real or it is not significant in the discourse. Private states and speech events may not be real in the discourse for several reasons: an example of one is when the private state or speech event is hypothetical. Private states or speech events that are not

Figure 3.4: **Objective speech event frame**

- **text anchor:** a pointer to the span of text that denotes the speech event.
- **source:** the speaker or writer.
- **target:** the target or topic of the speech event, i.e., the content of what is said.
- **properties:**
  - **implicit:** true, if the frame represents an an *implicit* speech event.
  - **insubstantial:** true, if the speech event is not substantial in the discourse.

significant are those that do not contain a significant portion of the contents (target) of the private state or speech event.

3.1.2 Objective Speech Event Frames

The **objective speech event frame** is used to mark speech events that do not express private states. They capture when material is attributed to some source, but is being presented objectively. An example of an objective speech event is "said" in sentence 3.7:

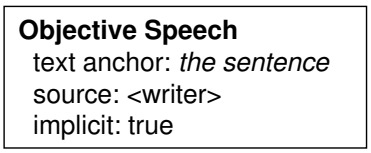
> (3.7) White House spokesman Jim Dyke said Miers' confirmation hearings are set to begin Nov. 7.

That Miers' confirmation hearings are to begin November 7 is presented as fact with White House spokesman Jim Dyke as the source of the information. The objective speech event frame, given in Figure 3.4, contains a subset of the attributes in private state frames. Although represented in the conceptualization, targets of objective speech events have not been annotated.

3.1.3 Agent Frames

The **agent frame** is used to mark noun phrases that refer to sources of private states and speech events. For example, agent frames would be created for "Democrats" in sentence 3.1, "LaBoon" in sentence 3.3, and "White House spokesman Jim Dyke" in sentence 3.7. As with the other frames, agents have *text anchor* and *source* attributes. For agents, the text

Figure 3.5: Speech event annotation for writer in sentence 3.8



anchor points to the span of text that denotes the noun phrase referring to the agent. The source of an agent frame is again a nested source. For example, the source for the agent frame for "White House spokesman Jim Dyke" is ⟨*writer*, *Jim Dyke*⟩. The nested source is composed of a list of alpha-numeric IDs. The IDs uniquely identify the agents in the nested source throughout the document. The agent frame associated with the first informative (e.g., non-pronomial) reference to a particular agent in the document includes an *ID* attribute to set up the document-specific agent-ID mapping.

3.1.4 Detailed Example

In this section, I describe the private state and speech event frames for a more complex example sentence.

> (3.8) The Foreign Ministry said Tuesday that it was "surprised, to put it mildly" by the U.S. State Department's criticism of Russia's human rights record, and objected in particular to the "odious" section on Chechnya.

For the writer of sentence 3.8, there is an objective speech event. The writer is simply presenting it as factual that the Foreign Ministry said what is reported, and the entire sentence is attributed to the writer. The objective speech event annotation for the writer is given in Figure 3.5.

The Foreign Ministry has several private state and speech event expressions in the sentence, which are given in Figure 3.6. The phrases "said" and "objected" are both speech events expressing private states. Both expressions are attributed to the Foreign Ministry by the writer of the sentence, and so have the source: ⟨*writer*, *Foreign Ministry*⟩. The phrase

"surprised, to put it mildly" is an explicit mention of a private state, attributed by the Foreign Ministry to itself. It has the source: ⟨*writer*, *Foreign Ministry*, *Foreign Ministry*⟩. There are two expressive subjective elements in the sentence: "to put it mildly" and "odious." These phrases are also directly attributed to the Foreign Ministry by the writer, and have the source: ⟨*writer*, *Foreign Ministry*⟩.

The speech event "said" is marked with a direct subjective frame rather than an objective speech frame because of the phrase "to put it mildly," which falls within scope of "said" and has the same source attribution. By itself, the phrase "to put it mildly" is subjective. As part of the larger phrase "surprised, to put it mildly," "to put it mildly" functions as an intensifier. Because "surprised" and "surprised, to put it mildly" are two different private states, the phase "to put it mildly" must also be included in the text anchor for the direct subjective frame. The intensity and expression intensity for "surprised, to put it mildly" is high. The intensity for "to put it mildly" is marked as medium. The intensity for "said" is also medium, but it has a neutral expression intensity.

As mentioned above, there is a direct subjective frame created for "objected," and an expressive subjective element created for "odious." Both "objected" and "odious" have an intensity of high. The word "odious" by itself expresses an intensively negative attitude. Because "odious" is included when evaluating the overall private state indicated by "objected," "objected" also has a high intensity. The expression intensity, which captures just the intensity of the text anchor, is marked as medium.

There is one more level of attribution in sentence 3.8. A direct subjective frame is created for the word "criticism" with source: ⟨*writer*, *Foreign Ministry*, *Foreign Ministry*, *US State Department*⟩. The word "criticism" is marked with an intensity and expression intensity of medium. It captures the negative attitude toward Russia. This annotation is given in Figure 3.7.

Figure 3.6: Private state annotations for the Foreign Ministry in sentence 3.8

Direct Subjective
text anchor: *said*
source: <writer, Foreign Ministry>
intensity: medium
expression intensity: neutral
attitude type: negative

Direct Subjective
text anchor: *surprised, to put it mildly*
source: <writer, Foreign Ministry, Foreign Ministry>
target: U.S. State Department
intensity: high
expression intensity: high
attitude type: negative

Expressive Subjective Element
text anchor: *to put it mildly*
source: <writer, Foreign Ministry>
intensity: medium
attitude type: negative

Direct Subjective
text anchor: *objected*
source: <writer, Foreign Ministry>
intensity: high
expression intensity: medium
attitude type: negative

Expressive Subjective Element
text anchor: *odious*
source: <writer, Foreign Ministry>
intensity: high
attitude type: negative

Figure 3.7: Private state annotation for U.S. State Department in sentence 3.8

Direct Subjective
text anchor: *criticism*
source: <writer, Foreign Ministry, Foreign Ministry, U.S. State Dept>
target: Russia
intensity: medium
expression intensity: medium
attitude type: negative

3.2 REALIZING THE CONCEPTUALIZATION

This section describes the production of the MPQA Corpus version 1.0, based on the above conceptual representation. I also describe my process for training annotators and the inter-annotator agreement studies that I conducted to validate that core aspects of the annotation scheme can be annotated reliably.

3.2.1 Corpus Production

To move from the conceptual representation above to a corpus with private state and speech event annotations, the first step was to select an annotation tool and to create a coding schema implementing the conceptual representation. GATE version 1.2 (Cunningham et al., 2002) was selected as the annotation tool because of its ease of use and because it provides stand-off annotations with byte-offsets into the original document. Claire Cardie created the initial coding schema; shortly thereafter, I assumed responsibility for changes and additional development of the schema.

3.2.2 Annotator Training

Annotator training begins with an annotator reading the coding manual (Wiebe, 2002), which presents the conceptual representation described in Section 3.1. After this, training proceeds in two stages. First, the annotator focuses on learning the conceptual representation. Then, the annotator learns how to create the annotations using GATE.

In the first stage of training, the annotator practices applying the annotation scheme to four to six training documents, using pencil and paper to mark the private state frames, objective speech frames, and their attributes. The training documents are not trivial. They are news articles from the world press, drawn from the same corpus of documents that the annotator will be annotating. When the annotation scheme was first being developed, these documents were studied and discussed in detail until consensus annotations were agreed upon that could be used as a gold standard. After annotating each training document, the annotator compares his or her annotations to the gold standard for the document. During

this time, the annotator is encouraged to ask questions, to discuss where his or her tags disagree with the gold standard, and to reread any portion of the conceptual representation that may not yet be perfectly clear.

After I judge that the annotator has a firm grasp of the conceptual representation and can consistently apply the scheme on paper, the annotator learns to apply the scheme using the annotation tool. First, the annotator reads the instructions (http://www.cs.pitt.edu/mpqa/opinion-annotations/gate-instructions) and works through a tutorial on performing the annotations using GATE. The annotator then practices by annotating two or three new documents using the annotation tool.

Using the conceptual representation and the additional training materials described above, I trained a total of six annotators over a four year period. Three of these annotators participated in the inter-annotator agreement study described in the next section.

3.2.3 Agreement Study

In this section, I test the general hypothesis that annotators can be trained to reliably annotate expressions of private states and their attributes. Specifically, I evaluate inter-coder agreement for the following key aspects of the conceptual representation:

1. Identifying text anchors for the combined set of *direct subjective* and *objective speech event* annotations
2. Identifying the text anchors for *expressive subjective element* annotations
3. Distinguishing between *direct subjective* and *objective speech event* annotations
4. Judging *intensity* and *expression intensity* attributes

For this study, three annotators (A, M, and S) independently annotated 13 documents with a total of 210 sentences. The articles are from a variety of topics and were selected so that 1/3 of the sentences are from news articles reporting on objective topics, 1/3 of the sentences are from news articles reporting on opinionated topics ("hot-topic" articles), and 1/3 of the sentences are from editorials.[1]

[1]The results presented in this section were reported in (Wilson and Wiebe, 2003; Wiebe, Wilson, and Cardie, 2005), with the exception of agreement for intensity judgments, which was reported in (Wilson, Wiebe, and Hwa, 2006).

In the instructions to the annotators, I asked them to rate the annotation difficulty of each article on a scale from 1 to 3, with 1 being the easiest and 3 being the most difficult. The annotators were not told which articles were about objective topics or which articles were editorials, only that they were being given a variety of different articles to annotate.

I hypothesized that the editorials would be the hardest to annotate and that the articles about objective topics would be the easiest. The ratings that the annotators assigned to the articles support this hypothesis. The annotators rated an average of 44% of the articles in the study as easy (rating 1) and 26% as difficult (rating 3). More importantly, they rated an average of 73% of the objective-topic articles as easy, and 89% of the editorials as difficult.

It makes intuitive sense that "hot-topic" articles would be more difficult to annotate than articles about objective topics and that editorials would be more difficult still. Editorials and "hot-topic" articles contain many more expressions of private states, requiring an annotator to make more judgments than he or she would have to for articles about objective topics.

3.2.3.1 Measuring Agreement for Text Anchors The first step in measuring agreement is to verify that annotators do indeed agree on which expressions should be marked. To illustrate this agreement problem, consider the words and phrases identified by annotators A and M in example 3.9. Text anchors for direct subjective frames are in bold; text anchors for expressive subjective elements are underlined.

(3.9)
A: We **applauded** this move because it was not only just, but it made us **begin to feel** that we, as Arabs, were an integral part of Israeli society.

M: We **applauded** this move because it was not only just, but it made us **begin to feel** that we, as Arabs, were an integral part of Israeli society.

In this sentence, the two annotators mostly agree on which expressions to annotate. Both annotators agree that "applauded" and "begin to feel" express private states and that "not only just" is an expressive subjective element. However, in addition to these text anchors, annotator M also marked the words "because" and "but" as expressive subjective elements. The annotators also did not completely agree about the extent of the expressive subjective

element beginning with "integral."

The annotations from example 3.9 illustrate two issues that need to be considered when measuring agreement for text anchors. First, how should agreement be defined for cases when annotators identify the same expression in the text, but differ in their marking of the expression boundaries? This occurred in 3.9 when annotator A identified the word "integral" and annotator M identified the overlapping phrase "integral part." The second question to address is which statistic is appropriate for measuring agreement between annotation sets that disagree with respect to the presence or absence of individual annotations.

Regarding the first issue, there was no attempt to define rules for boundary agreement in the annotation scheme or instructions, nor was boundary agreement stressed during training. For the purposes of this research, my collaborators and I believed that it was most important that annotators identify the same general expression, and that boundary agreement was secondary. Thus, for this agreement study, I consider overlapping text anchors, such as "integral" and "integral part," to be matches.

The second issue is that annotators will identify different sets of expressions as part of this task, and thus Cohen's Kappa (κ) (Cohen, 1960) is not an appropriate metric for evaluation. In example 3.9, the set of expressive subjective elements identified by annotator A is {"not only just", "integral"}. The set of expressive subjective elements identified by annotator M is {"because", "not only just", "but", "integral part"}. Cohen's κ is appropriate for tasks in which the annotators tag the same set of objects, for example, sense tags applied to a set of word instances. In contrast, measuring agreement for text anchors requires evaluating the intersection between the sets of expressions identified by the annotators. An appropriate evaluation metric for this is F-measure. When evaluating the performance of a system, F-measure is the harmonic mean of precision and recall. When evaluating two sets of annotations from different annotators, precision and recall can be calculated with either annotator standing in for the system, which in practice makes precision and recall interchangeable. If A and B are the sets of anchors annotated by annotators a and b, respectively, then the recall of a with respect to b ($recall(a\|b)$) is as follows:

$$recall(a\|b) = \frac{|A \text{ matching } B|}{|A|}$$

Table 3.1: Inter-annotator agreement: Expressive subjective elements

a	b	$recall(a\|\|b)$	$recall(b\|\|a)$	F-measure
A	M	0.76	0.72	0.74
A	S	0.68	0.81	0.74
M	S	0.59	0.74	0.66
average				0.71

Similarly, the recall of b with respect to a is:

$$recall(b||a) = \frac{|B \text{ matching } A|}{|B|}$$

3.2.3.2 Agreement for Expressive Subjective Element Text Anchors In the 210 sentences in the annotation study, the annotators A, M, and S respectively marked 311, 352 and 249 expressive subjective elements. Table 3.1 shows the pairwise agreement for these sets of annotations. For example, M agrees with 76% of the expressive subjective elements marked by A, and A agrees with 72% of the expressive subjective elements marked by M.

I hypothesized that the stronger the expression of subjectivity, the more likely the annotators are to agree. To test this hypothesis, I measure agreement for the expressive subjective elements rated with an intensity of *medium* or higher by at least one annotator. This excludes on average 29% of the expressive subjective elements, but results in an 8-point rise in average F-measure to 0.79. Similarly, when only expressive subjective elements rated as *high* or *extreme* by at least one annotator are considered, the average F-measure again increases another 8-points to 0.87, although with many more expressive subjective elements being excluded: 65% on average. Thus, annotators are more likely to agree when the expression of subjectivity is strong. Table 3.2 gives a sample of expressive subjective elements marked with high or extreme intensity by at least two annotators.

3.2.3.3 Agreement for Direct Subjective and Objective Speech Event Text Anchors This section measures agreement, collectively, for the text anchors of objective speech

Table 3.2: Sample of expressive subjective elements with high and extreme intensity

mother of terrorism
such a disadvantageous situation
will not be a game without risks
breeding terrorism
grown tremendously
menace
such animosity
throttling the voice
indulging in blood-shed and their lunaticism
ultimately the demon they have reared will eat up their own vitals
those digging graves for others, get engraved themselves
imperative for harmonious society
glorious
so exciting
disastrous consequences
could not have wished for a better situation
unconditionally and without delay
tainted with a significant degree of hypocrisy
in the lurch
floundering
the deeper truth
the Cold War stereotype
rare opportunity
would have been a joke

Table 3.3: Inter-annotator agreement: Explicitly-mentioned private states and speech events

a	b	$recall(a\|\|b)$	$recall(b\|\|a)$	F-measure
A	M	0.75	0.91	0.82
A	S	0.80	0.85	0.82
M	S	0.86	0.75	0.80
average				0.81

event and direct subjective frames. For ease of reference, in this section I refer to these frames collectively as **explicit frames**.[2] For the agreement measured in this section, frame type is ignored. The next section measures agreement between annotators in distinguishing objective speech events from direct subjective frames.

The three annotators, A, M, and S, respectively identified 338, 285, and 315 explicit frames in the data. Table 3.3 shows the agreement for these sets of annotations. The average F-measure for the text anchors of explicit frames is 0.81, which is 10-points higher than for expressive subjective elements. This shows that speech event and direct subjective frames are the easier frames to identify.

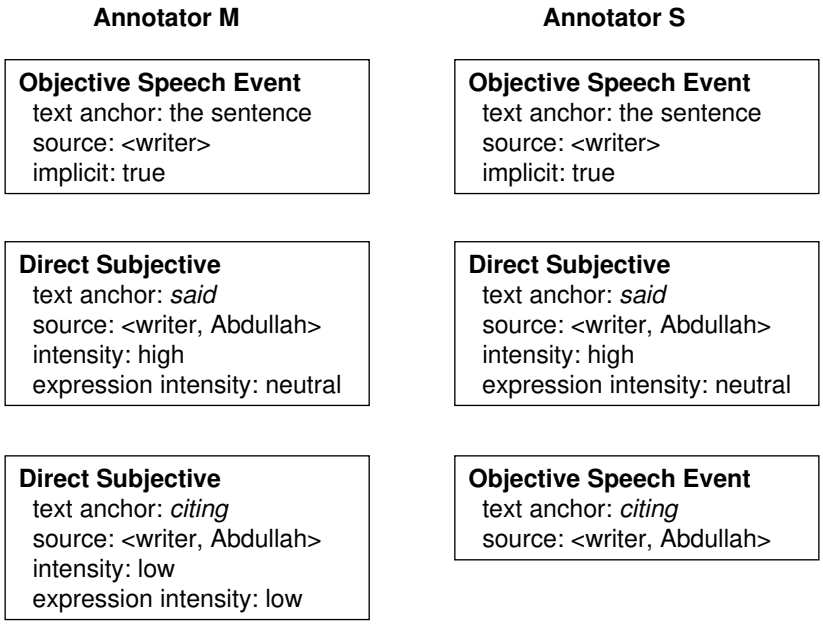
3.2.3.4 Agreement Distinguishing between Objective Speech Event and Direct Subjective Frames In this section, I focus on inter-rater agreement for judgments that reflect whether or not an opinion, emotion, or other private state is being expressed. I measure agreement for these judgments by considering how well the annotators agree in distinguishing between objective speech event frames and direct subjective frames. This distinction is considered to be a key aspect of the annotation scheme—a higher-level judgment of subjectivity versus objectivity than is typically made for individual expressive subjective elements.

For an example of the agreement I am measuring, consider sentence (3.10).

(3.10) "Those digging graves for others, get engraved themselves', he [Abdullah] said while

[2]Frames that are *implicit* with the source ⟨*writer*⟩ are excluded from this analysis. This is because the text anchors for the writer's implicit speech events are simply the entire sentence. The agreement for the text anchors of these speech events is trivially 100%.

Figure 3.8: Annotations in sentence 3.10 for annotators M and S



citing the example of Afghanistan.

Figure 3.8 gives the objective speech event frames and direct subjective frames identified by annotators M and S in sentence 3.10[3]. For this sentence, both annotators agree that there is an objective speech event frame for the writer and a direct subjective frame for Abdullah with the text anchor "said." They disagree, however, as to whether an objective speech event or a direct subjective frame should be marked for text anchor "citing." To measure agreement for distinguishing between objective speech event and direct subjective frames, I first match up the explicit frame annotations identified by both annotators (i.e., based on overlapping text anchors), including the frames for the writer's speech events. I then measure how well the annotators agree in their classification of that set of annotations as objective speech events or direct subjective frames.

Specifically, let $S1_{all}$ be the set of all objective speech event and direct subjective frames identified by annotator $A1$, and let $S2_{all}$ be the corresponding set of frames for annotator $A2$. Let $S1_{intersection}$ be all the frames in $S1_{all}$ such that there is a frame in $S2_{all}$ with an

[3]Some frame attributes not relevant for this agreement study have been omitted.

Table 3.4: Annotators A & M: Contingency table for objective speech event/direct subjective frame type agreement

		Tagger M	
		ObjectiveSpeech	*DirectSubjective*
Tagger A	*ObjectiveSpeech*	$n_{oo} = 181$	$n_{os} = 25$
	DirectSubjective	$n_{so} = 12$	$n_{ss} = 252$

overlapping text anchor. $S2_{intersection}$ is defined in the same way. The analysis in this section involves the frames $S1_{intersection}$ and $S2_{intersection}$. For each frame in $S1_{intersection}$, there is a matching frame in $S2_{intersection}$, and the two matching frames reference the same expression in the text. For each matching pair of frames, I am interested in determining whether the annotators agree on the type of frame: Is it an objective speech event or a direct subjective frame? Because the set of expressions being evaluated is the same, I use Cohen's κ to measure agreement.

Table 3.4 shows the contingency table for these judgments made by annotators A and M. n_{oo} is the number of frames the annotators agreed were objective speech events. n_{ss} is the number of frames the annotators agreed were direct subjective. n_{so} and n_{os} are their disagreements. The κ scores for all annotator pairs are given in Table 3.5. The average pairwise κ score is 0.81. Under Krippendorff's scale (Krippendorff, 1980), this allows for definite conclusions about the reliability of the annotations.

Table 3.5: Pairwise κ scores and overall percent agreement for objective speech event/direct subjective frame type judgments

	All Expressions		Borderline Removed		
	κ	agree	κ	agree	% removed
A & M	0.84	0.91	0.94	0.96	10
A & S	0.84	0.92	0.90	0.95	8
M & S	0.74	0.87	0.84	0.92	12

With many judgments that characterize natural language, one would expect that there are clear cases as well as borderline cases that are more difficult to judge. This seems to be the case with sentence 3.10 above. Both annotators agree that there is a strong private state being expressed by the speech event "said." But the speech event for "citing" is less clear. One annotator sees only an objective speech event. The other annotator sees a weak expression of a private state (the *intensity* and *expression intensity* ratings in the frame are *low*). Indeed, the agreement results provide evidence that there are borderline cases for objective versus subjective speech events. Consider the expressions referenced by the frames in $S1_{intersection}$ and $S2_{intersection}$. I consider an expression to be **borderline subjective** if (1) at least one annotator marked the expression with a direct subjective frame and (2) neither annotator characterized its intensity as being greater than low. For example, "citing" in sentence 3.10 is borderline subjective. In sentence 3.11 below, the expression "observed" is also borderline subjective, whereas the expression "would not like" is not. The frames identified by annotators M and S for sentence 3.11 are given in Figure 3.9.

(3.11) "The US authorities would not like to have it [Mexico] as a trading partner and, at the same time, close to OPEC," Lasserre observed.

Table 3.6 gives the contingency table for the judgments given in Table 3.5 but with the frames for the borderline subjective expressions removed. This removes, on average, only 10% of the expressions. When these are removed, the average pairwise κ climbs to 0.89.

3.2.3.5 Agreement for Sentences In this section, I use the annotators' low-level frame annotations to derive sentence-level judgments, and I measure agreement for those judgments.

Measuring agreement using higher-level summary judgments is informative for two reasons. First, objective speech event and direct subjective frames that were excluded from consideration in Section 3.2.3.4 because they were identified by only one annotator[4] may now be included. Second, having sentence-level judgments enables us to compare agreement for our annotations with previously published results (Bruce and Wiebe, 1999).

[4]Specifically, the frames that are not in the sets $S1_{intersection}$ and $S2_{intersection}$ were excluded.

Figure 3.9: Annotations in sentence 3.11 for annotators M and S

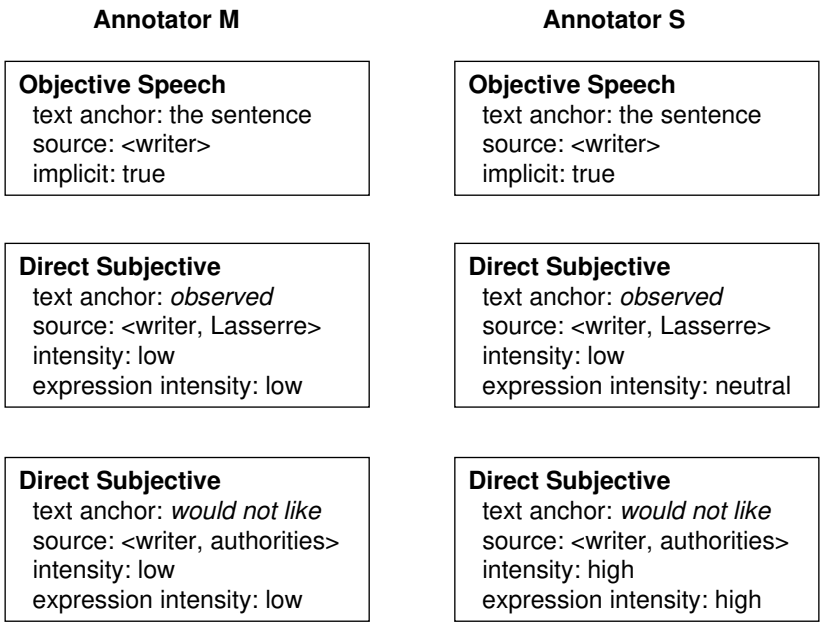


Table 3.6: Annotators A & M: Contingency table for objective speech event/direct subjective frame type agreement, borderline subjective frames removed

		Tagger M	
		ObjectiveSpeech	*DirectSubjective*
Tagger A	*ObjectiveSpeech*	$n_{oo} = 181$	$n_{os} = 8$
	DirectSubjective	$n_{so} = 11$	$n_{ss} = 224$

Table 3.7: Pairwise κ scores and overall percent agreement for sentence-level objective/subjective judgments

	All Sentences		Borderline Removed		
	κ	agree	κ	agree	% removed
A & M	0.75	0.89	0.87	0.95	11
A & S	0.84	0.94	0.92	0.97	8
M & S	0.72	0.88	0.83	0.93	13

The annotators' sentence-level judgments are defined in terms of their lower-level frame annotations as follows. First, I exclude the objective speech event and direct subjective frames that both annotators marked as *insubstantial*. Then, for each sentence, an annotator's judgment for that sentence is *subjective* if the annotator created one or more direct subjective frames in the sentence; otherwise, the judgment for the sentence is *objective*.

The pairwise agreement results for these derived sentence-level annotations are given in Table 3.7. The average pairwise κ for sentence-level agreement is 0.77, 8 points higher than the sentence-level agreement reported in (Bruce and Wiebe, 1999). This result suggests that adding detail to the annotation task may help annotators perform more reliably.

As with objective speech event versus direct subjective frame judgments, I again test whether removing borderline cases improves agreement. I define a sentence to be **borderline subjective** if (1) at least one annotator marked at least one direct subjective frame in the sentence, and (2) neither annotator marked a direct subjective frame with an *intensity* greater than low. When borderline subjective sentences are removed, on average only 11% of sentences, the average κ increases to 0.87.

3.2.3.6 Agreement for Intensity Judgments This section reports on the agreement for the intensity and expression intensity of private states and speech events, and for the intensity of expressive subjective elements. For the experiments presented later in Chapter 5, I merge the *high* and *extreme* intensity classes because of the rarity of the *extreme* class (only 2% of sentences in the corpus contain an annotation with extreme intensity). Thus, when

calculating agreement for the various intensity judgments, I also merge the *high* and *extreme* ratings, to mirror their treatment in the experiments.

Included in the judgment of intensity is a determination of whether a private state is being expressed at all. That is, when an annotator chooses to mark an expression as an objective speech event as opposed to a direct subjective annotation, the annotator is in essence making a judgment that intensity is *neutral*. Thus, to accurately measure agreement for intensity, I consider direct subjective annotations and objective speech annotations together. The value of the intensity for all objective speech events is neutral. For all objective speech events that are not implicit, expression intensity is also neutral.

The classes used for intensity judgments represent an ordinal scale; this raises the question of which agreement metric is appropriate for evaluating intensity. For the combined direct subjective and objective speech event annotations, the rating scale for both intensity and expression intensity is *neutral*, *low*, *medium*, and *high*. For expressive subjective elements, the rating scale for intensity is *low*, *medium*, and *high*. Agreement metrics such as Cohen's κ treat all disagreements equally, which is suitable for discrete classes. However, with the ordinal nature of the intensity judgments, not all disagreements are equal. For example, a disagreement about whether intensity is neutral or high is more severe than a disagreement about whether it is medium or high. Cohen's κ, therefore, is not a suitable metric.

There is an adaptation of Cohen's κ called weighted κ (Cohen, 1968), which is for use with ordinal data. Weighted κ assigns weights that allow for partial agreement. However, the weights are calculated based on the number of categories. The intensity scale used for direct subjective and speech event annotations is slightly different than the one for expressive subjective elements, which doesn't include the *neutral* class. This means that with weighted κ, the weights for expressive subjective elements will be different than the weights for direct subjective and speech event annotations. Because of this, weighted κ is also inappropriate.

The metric that I use for agreement for intensity judgments is Krippendorff's α (Krippendorff, 1980; Krippendorff, 2004). Like Cohen's κ, Krippendorff's α takes into account chance agreement between annotators, but it is more general. It can be used to calculate agreement for both discrete and ordinal judgments, and its method of weighting disagree-

ments does not depend on the number of categories. In its most general form, α is defined to be

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o is a measure of the observed disagreement and D_e is a measure of the disagreement that can be expected by chance. Krippendorff's α ranges between 0 and 1, with $\alpha = 1$ indicating perfect agreement and $\alpha = 0$ indicating agreement that is no better than chance.

With α, a distance metric is used to weight disagreements. Different distance metrics are used for different types of data. For intensity, the ratings map naturally to the scale [0,1,2,3], where 0 represents neutral and 3 represents high. Using this scale, I can use the distance metric that squares the difference between any two disagreements. Thus, the distance weight is 1 for any disagreement that differs by one (e.g., neutral-low), the distance weight is 4 for any disagreement that differs by two (e.g., neutral-medium), and the distance weight is 9 for any disagreement that differs by three (e.g., neutral-high).

I measure agreement for the intensity of the combined direct subjective and speech event annotations using the set of matching frames identified in Section 3.2.3.4 (the matching frames in $S1_{intersection}$ and $S2_{intersection}$). This is the same set of annotations that are used for calculating agreement for distinguishing between objective speech event and direct subjective frames. To measure expression-intensity agreement, I also use this set, with the exclusion of the matching frames that are marked with the *implicit* attribute.

Table 3.8 gives the pairwise α-agreement values for the intensity and expression intensity judgments of the combined direct subjective and objective speech event annotations. For comparison, the absolute percent agreement is also given. In interpreting α, Krippendorff (2004) suggests that values above 0.8 indicate strong reliability and values above 0.67 are sufficient for tentative conclusions. Using this scale, we see that the α scores for the intensity judgments of direct subjective and speech events are good.

For expressive subjective elements, I again identify the set of matching annotations that were marked by both annotators. Table 3.9 gives the pairwise α-agreement for the intensity of this set of expressive subjective elements, along with absolute percent agreement for comparison. Unlike the agreement for the intensity judgments of direct subjective and speech

Table 3.8: α-agreement and percent agreement for intensity judgments for the combined direct subjective and objective speech annotations

	Intensity		Expression Intensity	
Annotator Pair	α	%	α	%
A & M	0.79	0.73	0.76	0.66
A & S	0.81	0.75	0.76	0.63
M & S	0.76	0.76	0.73	0.59
average	0.79	0.75	0.75	0.62

Table 3.9: α-agreement and percent agreement for expressive subjective element intensity judgments

	Intensity	
Annotator Pair	α	%
A & M	0.40	0.49
A & S	0.52	0.56
M & S	0.46	0.54
average	0.46	0.53

event annotations, agreement for the intensity judgments of expressive subjective elements is not high. A look at the disagreements shows that many of them are influenced by differences in boundary judgments. Although annotations are considered matching as long as they have overlapping text spans, differences in boundaries can affect how intensity is judged. Example 3.12 below shows how the same subjective expression was judged by two annotators.

(3.12)
A: <*high*>imperative for harmonious society</>
M: <*medium*>imperative</> for <*medium*>harmonious</> society

Both annotators recognized that the above phrase is subjective. However, while the first annotator marked the entire phrase as a single expressive subjective element with high intensity, the second annotator marked particular words and smaller phrases as expressive subjective elements and judged the intensity of each separately.

A severe type of disagreement between annotators is a difference in intensity ordering, i.e., annotator A rating expression 1 as more intense than expression 2, and annotator B rating expression 2 as more intense than expression 1. Fortunately, there are few such disagreements. On average, only 5% of all possible pairings of matching annotations result in disagreements in the ordering of intensity.

3.2.3.7 Agreement for Intensity of Sentences As with subjective and objective sentence-level judgments in Section 3.2.3.5, sentence-level intensity judgments can be derived from the expression-level intensity judgments. In this section, I measure agreement for those judgments.

Evaluating intensity agreement at the sentence level is important for two reasons. First, annotations that were previously excluded from consideration because they were identified by only one annotator may now be included. Second, for the experiments in Chapter 5, the units of evaluation are sentences and clauses, and it is important to know what the agreement is for intensity judgments at this higher level.

I define an annotator's intensity judgment for a sentence as the highest intensity or expression-intensity rating of any annotation marked by that annotator in the sentence.

Table 3.10: α-agreement and percent agreement for sentence-level intensity judgments

	Intensity	
Annotator Pair	α	%
A & M	0.74	0.56
A & S	0.83	0.73
M & S	0.72	0.57
average	0.77	0.62

Pairwise agreement scores for sentence-level intensity judgments are given in Table 3.10. The average α-agreement for sentences is 0.77.

3.2.4 Attitude Types and Targets

Although attitude types and targets are part of the conceptual representation of private states, they were not comprehensively annotated. Annotator training focused on the core aspects of the annotation scheme (e.g., distinguishing between direct subjective frames and objective speech events), and initially annotators were told to mark the attitude type for direct subjective frames and expressive subjective elements only when they felt comfortable doing so (i.e., for those private states that they felt expressed a clear and unambiguous polarity).[5] Targets were marked even less often. In the instructions, annotators were told to mark targets only for direct subjective frames that were clearly expressing a positive or negative attitude and only when the targets were agents in the discourse.

In essence, the attitude type and target annotations in this first version of the corpus were exploratory annotations. They helped to give a better sense of the notion of attitude type, which is more complex than simple polarity. This in turn helped to motive two extensions to the conceptual representation. The first extension adds an attribute to both private state frames specifically to represent the polarity of the text anchor expression. This extension is described in Chapter 6. The second extension, presented in Chapter 7, revises and further develops the conceptual representation for the attitudes and targets of private states.

[5]Private state frames with an attitude type of *none* are actually those for which the annotators did not mark an attitude type.

3.3 RELATED WORK

The conceptual representation for private states (Wiebe, 2002; Wiebe, Wilson, and Cardie, 2005) that forms the foundation for the work in this dissertation grew out of the model developed by Wiebe (1990; 1994) for tracking point of view in narrative. That model was in turn based on work in literary theory and linguistics, in particular, work by Doležel (1973), Uspensky (1973), Kuroda (1973; 1976), Chatman (1978), Cohn (1978), Fodor (1979), and Banfield (1982). The nested levels of attribution in the conceptual representation were inspired by work on propositional attudes and belief spaces in artificial intelligence (Wilks and Bien, 1983; Asher, 1986; Rapaport, 1986) and linguistics (Fodor, 1979; Fauconnier, 1985).

Of the few annotation schemes proposed for marking opinions and affect in text, the one most similar to the conceptual representation for private states is the framework proposed by Appraisal Theory (Martin, 2000; White, 2002) for analyzing evaluation and stance in discourse. Appraisal Theory emerged from the field of systemic functional linguistics (see Halliday (1985/1994) and Martin (1992)). The Appraisal framework is composed of the concepts (or *systems* in the terminology of systemic functional linguistics): Affect, Judgement, Appreciation, Engagement, and Amplification. Affect, Judgement, and Appreciation represent different types of positive and negative attitudes. Engagement distinguishes various types of "intersubjective positioning" such as attribution and expectation. Amplification considers the force and focus of the attitudes being expressed.

Appraisal Theory is similar to the conceptual representation of private states in that it, too, is concerned with systematically identifying expressions of opinions and emotions in context, below the level of the sentence. Force, which is part of the concept of Amplification, is similar to intensity. However, the two annotation schemes focus on different things. The Appraisal framework primarily distinguishes different types of private state (e.g., affect versus judgement), which I will review in greater detail after presenting my extensions to the scheme for modelling the attitudes of private states (Chapter 7). In contrast to the conceptual representation for private states, Appraisal Theory does not distinguish the different ways that private states may be expressed (i.e., directly, or indirectly using expressive subjective

elements), and it does not include a representation for nested levels of attribution. Although Appraisal Theory has been applied to various tasks in text and discourse analysis, whether the key concepts that compose the Appraisal framework can be reliably annotated has not been empirically evaluated.

Besides Appraisal Theory, subjectivity annotation of text in context has also been performed by Yu and Hatzivassiloglou (2003), Bethard et al. (2004), Kim and Hovy (2004), Bruce and Wiebe (1999), and Wiebe et al. (2004). The annotation schemes used in Bruce and Wiebe (1999) and Wiebe et al. (2004) are earlier, less detailed versions of the conceptual representation of private states. The annotations in Bruce and Wiebe (1999) are sentence-level annotations; the annotations in Wiebe et al. (2004) mark only the texts spans of expressive subjective elements. In contrast to the detailed, expression-level annotations of the annotation scheme described in this chapter, the corpora developed by Yu and Hatzivassiloglou (2003), Bethard et al. (2004), and Kim and Hovy (2004) provide only sentence-level subjectivity and/or sentiment annotations.

3.4 CONCLUSIONS

In this Chapter, I described the conceptual representation for private states that forms the foundation for the work in this dissertation. I hypothesized that annotators could be trained to reliably annotate the core aspects of the representation, specifically, to identify the text anchors for the private state and speech event annotations, to distinguish between direct subjective and objective speech event annotations, and to judge the intensity of private states.

To test this hypothesis, I conducted an inter-annotator agreement study with three trained annotators. To measure agreement for text anchors, I used F-measure to evaluate the intersection of the sets of text anchors marked by two annotators. Average F-measure for marking text anchors for the combined set of direct subjective and objective speech event frames is 0.81, and average F-measure for text spans of expressive of expressive subjective elements is 0.71. To measure agreement for distinguishing between direct subjective and

objective speech event frames I used Cohen's κ, and for intensity judgments I used Krippendorff's α. For distinguishing between direct subjective and objective speech event frames, average pairwise κ is 0.81, and average pairwise α-agreement for intensity judgments, again for the combined set of direct subjective and objective speech event frames, is 0.79.

For both κ and α, scores of 0.67 and higher are considered sufficient for drawing conclusions about annotation reliability, with higher scores providing stronger evidence. Thus, annotators can reliably distinguish between expressions of direct subjectivity and objective speech events. Agreement for intensity judgments of these expressions is also good. However, there are no standards that have been proposed for interpreting different values of F-measure, which makes it difficult to draw conclusions about how reliably annotators can identify text anchors. Agreement for the text anchors of direct subjective and objective speech event frames seems good, although there is room for improvement. What is clear is that direct subjective and objective speech event frames are easier to identify than expressive subjective elements. Also, when text anchor disagreements are included and sentence-level agreement is calculated using interpretable metrics, agreement is good. Average pairwise κ-agreement for sentence subjectivity is 0.77, and average pairwise α-agreement for sentence intensity is also 0.77. Although this does not explicitly speak to the reliability of annotating text anchors, it does show that annotators agree on higher-level judgments, even if there is some disagreement about individual expressions.

4.0 OVERVIEW OF MACHINE LEARNING EXPERIMENTS

One of the primary goals of the machine learning experiments in this dissertation is to evaluate features for recognizing the intensity, polarity, and attitudes of private states. To be able to draw strong conclusions about the utility of features for a given task, two things are needed: (1) a good experimental baseline to serve as a point for comparison, and (2) an evaluation that measures the performance of the features across a range of different machine learning algorithms.

Baselines are needed to provide points of reference for judging whether the features being evaluated actually are proving useful, but what makes a good baseline? A good baseline is one that is informed by the current level of existing knowledge for a given task. This may be knowledge as basic as the distribution of classes or perhaps information about the types of features already known to be helpful. What makes a good baseline will vary depending on the task, and perhaps also on knowledge gleaned from earlier work. Essentially, the question that must be answered is whether the knowledge represented by the features being evaluated extends beyond the current level of existing knowledge, which should be represented by the baseline.

Evaluating features using different learning algorithms can reveal a great deal about the features' utility. When features give a good performance across several different algorithms, this is strong evidence that the features are robust and important for that particular task. Similarly, when features give a weak performance across multiple algorithms, this is strong evidence that the features are less useful. When features give a mixed performance, improving results for some algorithms but not others, this shows that the features are likely good for that task, but that they are less robust.

In the sections below, I briefly discuss the types of experimental baselines (Section 4.1 and

describe the machine learning algorithms (Section 4.2) that are used in the experiments in the chapters that follow. I also discuss the issue of tuning algorithm parameters in Section 4.3. I end this chapter by considering what the experiments in this dissertation taken together may reveal about the kinds of features and algorithms that would be most useful for fine-grained subjectivity analysis in general.

4.1 EXPERIMENTAL BASELINES

As previously mentioned, a good baseline is one that is informed by the current level of existing knowledge for a given task. When deciding on the baselines for the experiments in this dissertation, three basic types were considered: (1) baselines based on a simple classifier that would always make the same informed prediction, (2) baselines based on the set of words in the unit being classified (bag-of-words), and (3) baselines that utilize prior knowledge about subjective words obtained from a lexicon. Each of these types of baselines represents a different kind knowledge.

An example of a baseline classifier that always makes the same informed prediction is a classifier that always chooses the most frequent class based on the distribution of classes in the training data. This is a fairly common baseline strategy, and, although it is simple, it can make for a challenging baseline, especially if the class distribution is heavily skewed. A most-frequent-class baseline classifier is used for the experiments in intensity recognition because, for some of the experiments, the accuracy of this classifier was higher than for the classifier trained using bag-of-words. A second baseline is also used for these experiments: a baseline classifier that always chooses the midpoint on the intensity scale. The choice of these two baselines is discussed in more detail in Chapter 5.

For the experiments in polarity and attitude recognition, bag-of-words classifiers are used for the baselines. The set of words contained in the unit to be classified represents a very basic, easily accessible, and often important source of information for NLP tasks. This is certainly true for polarity and attitude recognition. For these tasks, the knowledge represented by the set of words is enough to outperform the simple strategy of choosing the

most frequent class. For the polarity classification experiments, a second, more challenging baseline is also used. This baseline combines bag-of-words features with knowledge about the prior polarity of words from a subjectivity lexicon.

4.2 ALGORITHMS

There were several considerations in choosing which supervised machine learning algorithms to use for the types of experiments planned for this work. Most important was that the algorithms represent a range of different types of machine learning: The greater the variety of algorithms used in an experiment, the stronger the conclusions that can be drawn about the features being evaluated. It also would be good if the algorithms that are chosen have been used successfully for a number of natural language processing tasks. Although the aim is not to determine which algorithms are best for a particular task, it still would be good to choose algorithms that can be expected to perform well. Finally, for practical purposes the algorithms should have implementations that are straightforward to run and configure in terms of feature sets and parameters.

The types of machine learning that I chose to use are rule learning, boosting, support vector machines (SVMs), and instance-based learning (k-nearest neighbor). For rule learning I use Ripper (Cohen, 1996), a classic rule-induction classification algorithm that has been used for any number of NLP tasks. One plus for using a rule learner like Ripper is that the output of the algorithm is human readable and easily to understand, which can help to give insights into the problem.

Boosting is a type of machine learning that seeks to combine many weak learners into one highly accurate classifier. An example of a weak learner is a single, simple, not-very-accurate categorization rule. The weak learners are trained in iterations, with each iteration or round adding a new weak learner to the final classifier. Whereas the classifier learned by rule learning is a cascading set of if-then rules that are applied in a particular order (analogous to a decision tree), the boosting classifier is a set of weighted rules that are combined irrespective of order. The boosting classifier that I use is BoosTexter AdaBoost.MH (Schapire and Singer,

2000).

SVMs or maximum-margin classifiers represent the data as a set of vectors in an n-dimentional space. The goal of the algorithm is then to find the hyperplane in that space that separates the positive examples from the negative with the widest margin. SVMs can be trained to perform both classification and regression. The SVM algorithms that I use are SVM-light and SVM-multiclass (Joachims, 1999). SVM-light is used for regression and binary-classification experiments. SVM-multiclass is used for experiments involving the classification of more than two categories.

Instance-based or memory-based learning is a very different type of learning from the above algorithms. Rule learning, boosting, and SVMs put all their "effort" into training, i.e., using the training instances to build a model that hopefully will generalize and be able to classify unseen instances. Instance-based learning, on the other hand, postpones any generalization from the training instances until a new instance needs to be classified. Training for instance-based learning is simply a matter of reading all the training instances into memory. To classify a new instance, the most similar instances are retrieved from memory and used to predict the class of the new instance. The instance-based learning algorithm that I use is TiMBL IB1 (Daelemans et al., 2003b), a k-nearest neighbor algorithm.

The above algorithms represent several different types of machine learning and all have been used successfully on a variety of NLP tasks. Not all algorithms are used in the experiments in each chapter. The experiments in intensity recognition (Chapter 5) use Ripper, BoosTexter, and SVM-light regression. All four algorithms are used in the contextual-polarity experiments (Chapter 6). The attitude recognition experiments (Chapter 8) use BoosTexter and SVM-light.

4.3 PARAMETER TUNING

An important issue to address when working with machine learning algorithms is the tuning of the algorithms' parameters. Each of the algorithms that I use in my experiments have multiple parameters that can be configured to change the algorithm's behavior. For exam-

ple, the number of rounds of boosting can be varied for BoosTexter, different kernels can be used for SVM-light, and the number of neighbors to consider can be varied for TiMBL. Research has shown that the performance of learning algorithms for NLP tasks can vary widely depending on their parameter settings, and that the optimal parameter settings can also vary depending on the set of features being evaluated (Daelemans et al., 2003a; Hoste, 2005). Although it is not the goal of this research to identify the optimal parameter configuration for each algorithm for each experiment and each set of features being evaluated, it is important to make a reasonable attempt to identify a good configuration for each algorithm.

For the machine learning experiments in the following chapters, parameter settings for a given algorithm are selected through preliminary, baseline experiments on a set of development data, which is separate from the test data used in the experiments. Specifically, select parameter settings for an algorithm are varied while repeatedly training and evaluating the algorithm on the development data. Whichever settings result in the best performance on the development data are then used in the experiments on the test data.

The above method of parameter tuning should find good configurations for the machine learning algorithms, but it will not necessarily find optimal configurations. This is because the parameter search is not exhaustive and only select parameters are varied as part of the tuning process. Because the classifiers learned are not necessarily optimal, it will not be possible to draw any strong conclusions about the relative performance of one algorithm compared to another. If one algorithm performs worse than another, it is always possible that a different parameter configuration exists that would give an equal performance, but that this configuration was out of the bounds of the search.

4.4 DRAWING LARGER CONCLUSIONS

In the same way that evaluating features over multiple algorithms can lead to conclusions about the utility of features for a particular task, taking a broader perspective and considering the results of experiments from task to task can lead to conclusions about what features and algorithms may be good in general for fine-grained subjectivity analysis. This knowledge

could then be used to inform solutions for new problems in subjectivity analysis.

Although the exact features being explored vary from task to task, there are some similarities in the *kinds* of features being used, for example, features associated with sets of subjectivity clues are used for each task, even if the actual sets differ that are defined. Also, all tasks include features that capture syntactic information. If a certain kind of feature generally performs well across the different tasks, this is good evidence that that type of feature is generally useful for fine-grained subjectivity analysis.

Although it may not be possible to draw strong conclusions about whether a given algorithm is or is not the best one for a particular task, looking at how an algorithm performs across the various tasks should again be revealing. Such an analysis should suggest what kinds of features the various algorithm can best exploit and also how appropriate the different algorithms are for tasks that are more or less fine grained.

5.0 RECOGNIZING STRONG AND WEAK OPINION CLAUSES

In this chapter, I investigate the automatic classification of the **intensity** of clauses and sentences. Recognizing intensity includes not only discriminating between private states of different intensity levels, but also detecting the absence of private states. Thus, the intensity classification task subsumes the task of classifying language as subjective versus objective.

My approach to this task is to use supervised machine learning techniques to train classifiers to predict intensity. The learning algorithms use a large lexicon of **subjectivity clues**, summarized in Section 5.3. Subjectivity clues are words and phrases that may be used to express private states. The clues in the lexicon are diverse. Many were learned automatically or collected from manual resources in previous studies of subjective language. The lexicon also contains new syntactic clues, which were developed for this research by Rebecca Hwa. People use a staggering variety of words and phrases to express opinions. With the new syntactic clues, one goal is to capture common dependencies between words that may be relevant for recognizing intensity, such as intensifying adverbs modifying adjectives (e.g., *quite good* and *very bad*).

For the learning algorithms to take full advantage of the subjectivity clues in the lexicon, there are two major challenges that must be addressed. One is the sheer volume of clues; the other is that many of the words and phrases in the lexicon occur very infrequently. This raises the question of how best to organize the clues in the lexicon into features for the learning algorithms. The approach I use is to organize the clues into sets and to create one feature per set. Section 5.4 describes the two different methods I use for organizing clues into sets, and how features for the learning algorithms are defined based on these sets.

For both training and testing I use the MPQA Corpus version 1.0, with the detailed private state annotations described in the previous chapter. These annotations are used to

define the intensity of the sentences and clauses for the experiments, which are presented in Section 5.5. Through the experiments, I show that many clues of subjective language, including the new syntactic clues and those from the literature, can be adapted to the task of intensity recognition. The experiments further show that the best results for intensity classification are achieved when the widest variety of clues is used.

This chapter is organized as follows. Section 5.1 briefly describes the division of the MPQA Corpus into datasets for experiments. In Section 5.2, I consider what determines and changes the intensity of expressions. Section 5.3 describes the lexicon of subjectivity clues used for the intensity classification experiments, and Section 5.4 describes the feature organizations that are used. In Section 5.5, I present the experiments and results of intensity classification, and I conclude the chapter in Section 5.7.

5.1 DATASETS

For the experiments in this chapter, the documents in the MPQA Corpus 1.0 are divided into two datasets. The first dataset (66 documents/1,344 sentences) is a development set, used for data exploration, feature development, and parameter tuning. The second dataset (469 documents/9,313 sentences) is an evaluation set, used to identify and evaluate the new syntactic clues presented below in Section 5.3.2 and in the experiments in Section 5.5. The sentences in the evaluation set are further divided into 10 folds, which are used to define training and testing sets for cross validation.

5.2 EXPLORING INTENSITY

An examination of the portion of annotated data held out for development shows not only that an extreme variety of expressions have been marked, but that higher-intensity private states in particular are expressed in many different ways. Table 5.1 gives a sample of some subjective expressions with high and extreme intensity. Of course there are obvious words

that almost always express more intense private states, such as "exciting" and "hate." These are easy to list, as are some obvious modifications that increase or decrease their intensity: "**very** exciting," "**really** hate," and "**don't** hate." However, it is unlikely that expressions like "powder keg," "freak show," "pre-historic," and "tyrannical" readily come to mind, all of which are marked in the MPQA Corpus.

Higher-intensity expressions often contain words that are very infrequent. For example, the words "pre-historic," "tyrannical," and "lunaticism" each appear only once in the corpus. Because subjective words are often less frequent (Wiebe et al., 2004), it is important to have knowledge of patterns like "expressed <direct-object>," which can generalize to many different phrases, such as "expressed hope," "expressed concern," "expressed gratitude," and "expressed some understanding." Collocations like "at all" add punch to an expression, as in, "at all costs" and "not true at all." There are also syntactic modifications and syntactic patterns that have subjective force. In addition to those patterns that merely intensify a subjective word, for example "very <ADJECTIVE>", there are patterns that have a cumulative effect on intensity: "justice and freedom," "terrorist and extremist," "violence and intimidation," "exaggerations and fabrications," and "disdain and wrath." The clues used later in the intensity classification experiments contain examples of all these kinds of subjective phenomena.

Sentences in which private states are expressed are often complex, with subjective expressions of differing intensities being expressed by perhaps two or more agents. This is the case in sentence 5.1 below.

(5.1) President Mohammad Khatami of Iran, whose attempt at reforms have gotten American <low>support</>, <high>accused</> the United States of "<high>warmongering</>."

In this sentence, there is low-intensity support being expressed by the United States, as well as high-intensity negative accusations coming from Khatami. In the MPQA Corpus, 31% of sentences are made up of clauses that differ in intensity by two or more intensity ratings. This highlights the need to identify opinions at the clause level, as I do in the experiments in this chapter.

Table 5.1: Sample of subjective expressions with high and extreme intensity ratings

victory of justice and freedom	such a disadvantageous situation
will not be a game without risk	breeding terrorism
grown tremendously	menace
such animosity	not true at all
throttling the voice	imperative for harmonious society
tainted with a significant degree of hypocrisy	power at all costs
so exciting	disastrous consequences
violence and intimidation	did not exactly cover himself in glory
could not have wished for a better situation	exalted
freak show	the embodiment of two-sided justice
if you're not with us, you're against us	appalling
vehemently denied	very definitely
everything good and nice	diametrically opposed
under no circumstances	shameful mum
purposes of intimidation and exaggeration	justice-seeking cries
should be an eye opener for the whole world	powder keg
most fraudulent, terrorist and extremist	enthusiastically asked
number one democracy	hate
apocalyptic savagery	gross misstatement
odious	increasingly tyrannical
indulging in blood-shed and their lunaticism	surprised, to put it mildly
glorious	disdain and wrath
many absurdities, exaggerations, and fabrications	great fanfare
take justice to pre-historic times	unconditionally and without delay
so conservative that it makes Pat Buchanan look vegetarian	
those digging graves for others, get engraved themselves	
lost the reputation of commitment to principles of human justice	
ultimately the demon they have reared will eat up their own vitals	

5.3 SUBJECTIVITY CLUES

In this section, I describe the knowledge that I use for automatic intensity classification, namely a broad collection of **subjectivity clues**. Subjectivity clues are words and phrases that may be used to express private states. In other words, they have subjective usages, although they may have objective usages as well. The subjectivity clues that I use include words and phrases from an established subjectivity lexicon and new syntactic clues that are correlated with subjective language.

I begin by reviewing the wide variety of clues in the established subjectivity lexicon. I then describe the collection of new syntactic clues, which were identified for this research by Rebecca Hwa.

5.3.1 Previously Established Types of Clues

Previous work in subjectivity analysis has led to the development of a large lexicon of subjectivity clues. I refer to the clues in this lexicon as PREV clues. The PREV clues include words and phrases culled from manually developed resources and learned from annotated and unannotated data. An interesting aspect of the set of PREV clues is that, because of the wide variety of sources from which they were compiled, the lexicon is quite varied and is not limited to a fixed word list or to words of a particular part of speech.

The clues from manually developed resources include:

- Verbs of judgment (e.g., *commend*, *reprove*, *vilify*), desire (e.g., *fancy*, *pine*, *want*), and psych (e.g., *dread*, *love*, *vex*) from Levin's (1993) English verb classes.
- Words and phrases culled from Ballmer and Brennenstuhl's (1981) speech act verb classes (e.g., *advocate*, *grumble about*, *vow*).
- Verbs and adjectives listed in FrameNet (Baker, Fillmore, and Lowe, 1998) with frame element *experiencer*. These include words from the Emotion_active (e.g., *fuss*, *worry*), Emotion_directed (e.g., *pleased*, *upset*), Emotion_heat (e.g., *burn*, *seethe*), Experiencer_obj (e.g., *embarrass*, *thrill*), Experiencer_subj (e.g., *dislike*, *sympathize*), and Perception_body (e.g., *ache*, *tickle*) frames.

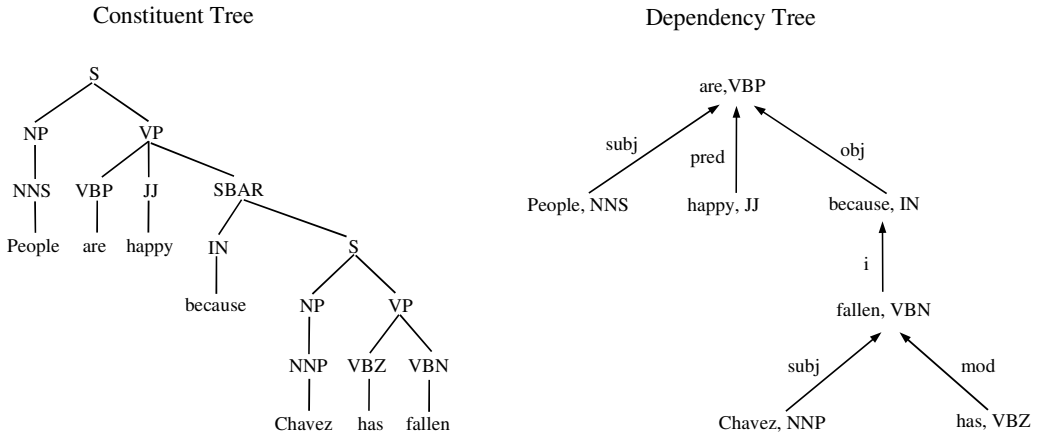
- Adjectives manually annotated for polarity from (Hatzivassiloglou and McKeown, 1997). The list of Positive adjectives includes the words *appealing*, *brilliant*, *luxurious*, and *nifty*. Included in the list of negative adjectives are the words *bizarre*, *dismal*, *hypocritical*, and *tedious*.
- Subjectivity clues listed in (Wiebe, 1990) (e.g., *absurdly*, *funny*, *grin*, *stench*, *truly*, *wonder*).

Clues learned from annotated data include distributionally similar adjectives and verbs, and n-grams from (Wiebe et al., 2004). The adjectives and verbs were learned from Wall Street Journal (WSJ) data using Dekang Lin's (1998) method for clustering words according to their distributional similarity. The seed words for this process were the adjectives and verbs in editorials and other opinion-piece articles. The n-gram clues were learned from WSJ data annotated for subjective expressions. They range from 2 to 4 words in length. Some examples of 3-grams are *worst of all*, *of the century*, and *do something about*. Examples of 4-grams are *on the other hand* and *price you have to*.

From unannotated data, extraction patterns and subjective nouns were learned using two different bootstrapping algorithms and a set of seed words (Riloff and Wiebe, 2003; Riloff, Wiebe, and Wilson, 2003). Extraction patterns are lexico-syntactic patterns typically used by information extraction systems to identify relevant information. For example, the pattern *<subject> was hired* would apply to sentences that contain the verb *hired* in the passive voice and would extract the subject as the hiree. In (Riloff and Wiebe, 2003), AutoSlog-TS, an algorithm for automatically generating extraction patterns, is used to find extraction patterns that are correlated with subjectivity. An example of a subjective extraction pattern is *<subj> dealt blow*, which matches phrases like "the mistake dealt a stiff blow to his pride." In (Riloff, Wiebe, and Wilson, 2003), the Meta-Bootstrapping (Riloff and Jones, 1999) and Basilisk (Thelen and Riloff, 2002) bootstrapping algorithms were used to learn sets of subjective nouns.

Finally, although not explicitly part of the lexicon, low-frequency words, which are informative for subjectivity recognition and require no training to identify (Wiebe et al., 2004), are also used as clues. A word is considered to be low frequency if it appears ≤ 3 times in the document containing it plus a 1-million word corpus of news articles. In addition, we use n-

Figure 5.1: The constituent and dependency parse trees for the sentence: *People are happy because Chavez has fallen*



gram clues from (Wiebe et al., 2004) that have fillers matching low-frequency words. When these clues were learned, the fillers matched low frequency words in the training data. When used during testing, the fillers are matched against low-frequency words in the test data. Examples of such n-grams are *<LowFreq-verb> and <LowFreq-verb>*, matching the phrases *bleat and bore* and *womanize and booze*, and *so <LowFreq-adj>*, matching the phrases *so enthusiastic* and *so cumbersome*.

Most of the above clues were collected as part of the work reported in (Riloff, Wiebe, and Wilson, 2003).

5.3.2 Syntax Clues

The new syntactic clues (SYNTAX clues) are developed by using a mostly-supervised learning procedure. The training data is based on both the annotations in the MPQA Corpus and a large unannotated corpus of automatically identified subjective and objective sentences from (Riloff and Wiebe, 2003). The procedure for learning the SYNTAX clues consists of three steps.

First, the training sentences in the MPQA corpus are parsed with a broad-coverage lexicalized English parser (Collins, 1997). The output constituent trees are automatically

converted into their dependency representations (Hwa and Lopez, 2004). In a dependency representation, every node in the tree structure is a surface word (i.e., there are no abstract nodes such as NP or VP), but each word may have additional attributes such as its part-of-speech (POS) tag. The parent word is known as the *head*, and its children are its *modifiers*. The edge between a parent and a child node specifies the grammatical relationship between the two words (e.g., *subj*, *obj*, and *adj*). Figure 5.1 shows the dependency parse tree for a sentence, along with the corresponding constituent representation, for comparison. For this study, 48 POS tags and 24 grammatical relationships are used.

Next, for each word in every dependency parse tree, all possible syntactic clues are exhaustively generated. There are five classes of syntactic clues. In addition, for each of the five classes, clues are generated that include specific words (indicated with **lex**) as well as less specific variants that back off to only POS tags (indicated with **backoff**).

root

root-lex(w, t): word w with POS tag t is the root of a dependency tree (i.e., the main verb of the sentence).

root-backoff(t): a word with POS tag t is the root of a dependency tree.

leaf

leaf-lex(w, t): word w with POS tag t is a leaf in a dependency tree (i.e., it has no modifiers).

leaf-backoff(t): a word with POS tag t is a leaf in a dependency tree

node

node-lex(w, t): word w with POS tag t.

node-backoff(t): a word with POS tag t.

bilex

bilex-lex(w, t, r, w_c, t_c): word w with POS tag t is modified by word w_c with POS tag t_c, and the grammatical relationship between them is r.

bilex-backoff(t, r, t_c): a word with POS tag t is modified by a word with POS tag t_c, and the grammatical relationship between them is r.

allkids

allkids-lex$(w, t, r_1, w_1, t_1, \ldots, r_n, w_n, t_n)$: word w with POS tag t has n children. Each

child word w_i has POS tag t_i and modifies w with grammatical relationship r_i, where $1 \leq i \leq n$.

allkids-backoff$(t, r_1, t_1, \ldots, r_n, t_n)$: a word with POS tag t has n children. The i^{th} child word has POS tag t_i and modifies the parent word with grammatical relationship r_i.

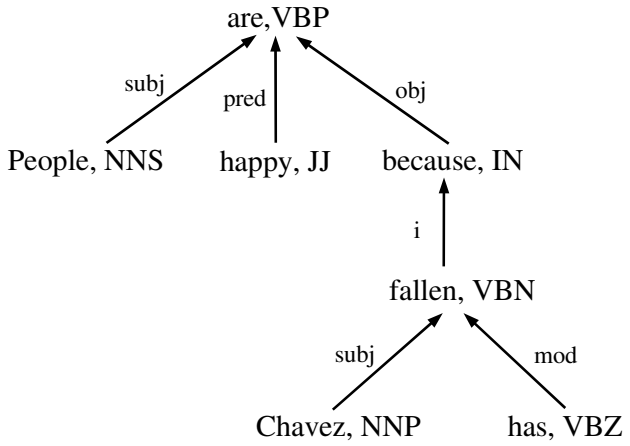
One thing that can determine the intensity of a private state being expressed is the patterning of a word together with its modifiers. For example, in the phrase *really quite nice*, the adverbial modifiers "really" and "quite" are working to intensify the positive evaluation expressed by "nice." With the *allkids* clues, the aim was to try to capture these types of patterns. One problem with the *allkids* clues, however, is the sparsity of their occurrences. This led to the inclusion of the *bilex* clues, which focus on the patterning found between a word and just one of its modifiers.

Examples of the different classes of syntactic clues are given in Figure 5.2. The top of Figure 5.2 gives the dependency parse tree for the sentence: *People are happy because Chavez has fallen.* The bottom half of the figure lists the potential syntactic-lex clues that would be generated for the sentence.

In a last step, the potential SYNTAX are evaluated to determine which clues to retain for the later experiments. A clue is considered to be *potentially useful* if more than x% of its occurrences are in subjective expressions in the training data, where x is a parameter tuned on the development set. For this work, $x = 70\%$ was chosen. Potentially useful clues are further categorized into one of three **reliability levels**. First, a clue is considered **highly reliable** if it occurs five or more times in the training data. For those that occur fewer than five times, their reliability is checked on the larger corpus of automatically identified subjective and objective sentences. Clues that do not occur in the larger unannotated corpus are considered **not very reliable**. Clues that occur in the subjective set at least y times more than in the objective set are considered **somewhat reliable**. The parameter y is tuned on the development set and is set to 4 in the experiments in this chapter. The remaining clues are rejected as not useful.

After filtering the potential syntax clues, 16,168 are retained on average per fold: 6.1% highly reliable, 42.9% somewhat reliable, and 51% not very reliable. Table 5.2 gives the distribution of clues based on type and reliability level. Table 5.3 gives a few examples of

Figure 5.2: Dependency parse tree and potential syntactic-lex clues generated from the tree for the sentence: *People are happy because Chavez has fallen*



root(are,VBP)
leaf(people,NNS)
leaf(happy,JJ)
leaf(chavez,NNP)
leaf(has,VBZ)
node(are,VBP)
node(people,NNS)
node(happy,JJ)
node(because,IN)
node(fallen,VBN)
node(chavez,NNP)
node(has,VBZ)
bilex(are,VBP,people,NNS,subj)
bilex(are,VBP,happy,JJ,pred)
bilex(are,VBP,because,IN,obj)
bilex(because,IN,fallen,VBN,i)
bilex(fallen,VBN,chavez,NNP,subj)
bilex(fallen,VBN,has,VBZ,mod)
allkids(are,VBP,people,NNS,subj,happy,JJ,pred,because,IN,pred)
allkids(because,IN,fallen,VBN,i)
allkids(fallen,VBN,chavez,NNP,subj,has,VBZ,mod)

Table 5.2: Distribution of retained syntax clues by type and reliability level

	Reliability Level		
Type	highly reliable	somewhat reliable	not very reliable
root	0.2	0.6	0.6
leaf	0.6	2.5	2.1
node	2.1	5.9	4.0
bilex	3.1	32.8	41.8
allkids	0.2	1.2	2.5

Values in the table are percentages.

allkids-backoff clues from the different reliability levels.

5.4 FEATURE ORGANIZATION

The large number of PREV and SYNTAX clues raises the question of how they should best be organized into features for intensity classification. A feature representation in which each clue is treated as a separate feature was tried, but this gave poor results. A likely reason for this is that so many of the individual clues are low frequency. Of the PREV clues with instances in the corpus, 32% only occur once and an additional 16% occur twice. With the SYNTAX clues, a full 57% have a frequency of one. Instead of treating each clue as a separate feature, I adopt the strategy of aggregating clues into sets and creating one feature for each set (Cohen, 1996; Wiebe, McKeever, and Bruce, 1998). The value of each feature is the number of instances in the sentence or clause of all the members of the set. The motivation for this type of organization is twofold. First, it increases the probability that a feature in the test set will have been observed in the training data: Even if a clue in the test set did not appear in the training data, other members of that clue's set may have appeared in the training data. Second, because clues are aggregated, feature frequencies are higher. I experiment with two strategies for aggregating clues into sets: organizing clues by their type and organizing clues by their intensity.

Table 5.3: Examples of *allkids-backoff* clues from different reliability levels and the instances that they match in the corpus

highly reliable	
CC,RB,mod,JJ,conj,JJ,conj	very precious *and* (very) sophisticated
	awfully grave *and* pressing
	only natural *and* rational
	quite neat *and* tidy
	thoroughly disgraceful *and* unacceptable
VB,DT,subj,JJ,pred	this *was* effective
	this *was* essential
	this *is* crazy
	those (who want to devalue) *are* shameless
	this *is* (no) different
somewhat reliable	
CC,JJR,conj,NN,conj,NN,conj	better governance *and* democracy
	greater speed *and* strength
WRB,JJ,adj,VB,i	*how* good (they) were
	how long (it can still) justify
	(no matter) *how* cynical (this may) appear
not very reliable	
VB,MD,mod,RP,mod,NN,obj,PREP,p	would *turn* back (the) clock on
WRB,NN,amod,VB,i	*where* (the) hell (it) is

For the instances, the word being modified is in italics, and words that are not its direct modifiers are in parentheses.

5.4.1 Organizing Clues by Type

To organize clues by their type, I define 29 sets for the PREV clues and 15 sets for the SYNTAX clues. The sets created for the PREV clues reflect how the clues were presented in the original research. For example, there are three sets created for the three classes of Levin (1993) verbs, and there are 2 sets created for the polar adjectives from (Hatzivassiloglou and McKeown, 1997), one for the positive adjectives and one for the negative adjectives. The SYNTAX clues are aggregated into sets based on the class of clue and its reliability level. For example, highly-reliable *bilex* clues form one set; somewhat-reliable *node* clues form another set.

In the experiments below, when features are used that correspond to sets of clues organized by type, they are referred to as TYPE features.

5.4.2 Organizing Clues by Intensity

Although the sets of subjectivity clues being used were selected because of their correlation with subjective language, they are not necessarily geared to discriminate between subjective language of differing intensities. Also, the groupings of clues into sets was not done with intensity in mind. I hypothesized that a feature organization that takes into consideration the potential intensity of clues would be better for intensity classification.

To adapt the clues for intensity classification, I use the annotations in the training data to filter the clues and organize them into four new sets, one for each intensity rating. Clues are placed into sets based on intensity as follows. For each clue c and intensity rating s, calculate $P(intensity(c) = s)$, the probability of c being in a subjective expression with intensity s. For $s = neutral$, this is the probability of c being in the text span of an objective speech event, in the text span of a direct subjective annotation with neutral expression-intensity, or in no annotation at all. Then, if $P(intensity(c) = s) \geq T(s)$, where $T(s)$ is the threshold determined for intensity s, place c in the set of clues with intensity s. In our experiments, we set $T(s) = P(intensity(word) = s) + 0.25$ or 0.95 if $P(intensity(word) = s) + 0.25 \geq 1$. $P(intensity(word) = s)$ is the probability of any given word being in a subjective expression with intensity s. The value 0.25 was determined using experiments on the development set.

Note that with this method of organizing clues into sets, it is possible for a clue to be in more than one set.

In the experiments below, when features are used that correspond to sets of clues organized by intensity, they are referred to as INTENSITY features.

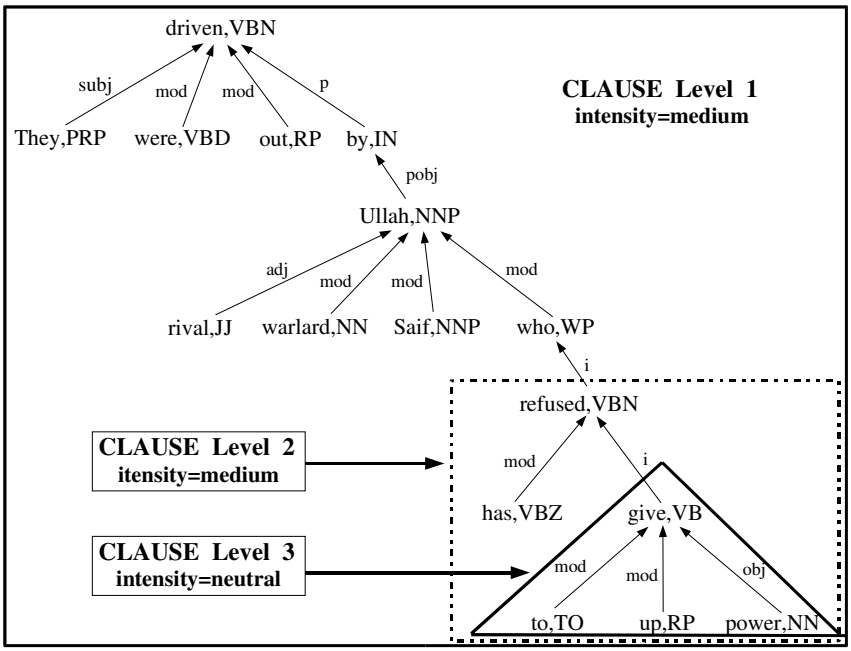
5.5 EXPERIMENTS IN INTENSITY CLASSIFICATION

It is important to classify the intensity of clauses as well as sentences, but pinpointing subjectivity at deeper levels can be challenging because there is less information to use for classification. To study the feasibility of automatically classifying clauses by their intensity, I conducted a suite of experiments in which an intensity classifier is trained based on the features previously described. My goal was to confirm three hypotheses. First, it is possible to classify the intensity of clauses, for those that are deeply nested as well as those at the sentence level. Second, classifying the intensity of subjectivity depends on a wide variety of features, including both lexical and syntactic clues. Third, a feature organization based on intensity is beneficial.

To test these hypotheses, I performed the experiments under different settings, varying four factors: (1) the learning algorithm used to train the classifiers, (2) the depth of the clauses to be classified, (3) the types of clues used, and (4) the feature organization (TYPE versus INTENSITY). The machine learning algorithms used in these experiments are BoosTexter AdaBoost.MH (Schapire and Singer, 2000), Ripper (Cohen, 1996), and SVM-light (support vector regression) (Joachims, 1999). To obtain the ordinal intensity classes for SVM-light, the predictions of the algorithm are discretized.

In the sections below, I first describe how the clauses for the experiments are determined, and how the gold-standard intensity classes are defined for sentences and clauses. I then describe the training-testing setup used for the experiments, followed by the experimental results.

Figure 5.3: Dependency parse tree and clauses for the sentence: *They were driven out by rival warlord Saif Ullah, who has refused to give up power*



5.5.1 Determining Clauses and Defining the Gold Standard

Clauses were chosen as the unit of evaluation because they can be determined automatically, and because they give different levels of nesting to vary in the experiments. Clauses are determined based on the non-leaf verbs in the parse tree of a sentence, parsed using the Collins parser and converted to the dependency representation as described earlier in Section 5.3.2. For example, sentence 5.2 has three clauses, corresponding to the verbs "driven," "refused," and "give."

(5.2) They were driven out by rival warlord Saif Ullah, who has refused to give up power.

The parse tree for sentence 5.2 is given in Figure 5.3. The clause defined for "driven" (level 1) is the entire sentence; the clause for "refused" (level 2) is "has refused to give up power"; and the clause for "give" (level 3) is "to give up power." Determining clauses in this way results in 9,817 level-2 clauses, 6,264 level-3 clauses, and 2,992 level-4 clauses in the experiment dataset.

The gold standard intensity ratings of sentences and clauses are based on the individual expression annotations; the intensity of a sentence or clause is defined to be the highest intensity rating of any expression in that sentence or clause. For example, in sentence 5.2, "refused" is the annotation with the highest intensity in the sentence. It was marked as a direct subjective annotation with medium intensity. Thus, the level-one clause (the entire sentence, headed by "driven") and the level-two clause (headed by "refused") both have a gold-standard intensity of medium. However, the gold-standard intensity for the level-three clause (headed by "give") is neutral, because the annotation for refused lies outside of the clause and there are no other annotations within the clause.

5.5.2 Experimental Setup

In setting up experiments for classifying nested clauses, there is a choice to be made for training: 1) either clauses from the same nested level may be used for training, or 2) clauses from a different level may be used for training. In the experiments in this paper, the training examples are always entire sentences, regardless of the clause level being classified during testing. Experimental results showed that this configuration is better than training on clauses from the same level. I believe this is because whole sentences contain more information.

5.5.3 Classification Results

All results reported are averages over 10-fold cross-validation experiments using the 9,313 sentences from the experiment dataset. Significance is measured using a 1-tailed t-test. For each experiment, both mean-squared error and classification accuracy are given. Although raw accuracy is important, it treats a misclassification that is off by 1 the same as a misclassification that is off by 3. As with disagreements in annotator intensity judgments, treating all intensity misclassifications equally doesn't reflect the ordinal nature of the intensity classes. Mean-squared error captures this distinction, and, for this task, it is perhaps more important than accuracy as a metric for evaluation. If t_i is the true intensity of sentence i, and p_i is the predicted intensity of sentence i,

$$\text{mean-squared error (MSE)} = \frac{1}{n}\sum_{i}^{n}(t_i - p_i)^2$$

where n is the number of sentences or clauses being classified. Note that the distance metric used in the α-agreement score back in Chapter 3 when measuring agreement for intensity judgments (Section 3.2.3.6) is the same as mean-squared error.

Table 5.4 gives the baselines and the results for experiments using all clues (PREV and TYPE) as well as experiments using bag-of-words (BAG). The question of what to use for a baseline is not straightforward. A common strategy is to use a baseline classifier that always chooses the most frequent class. However, the most frequent class for sentences is medium, which is different than the most frequent class for nested clauses, neutral. Thus, in Table 5.4 I chose to give both baselines, one for a classifier that always chooses neutral, and one for a classifier that always chooses medium. Note that there is quite a difference between the performance of the baselines with respect to mean-squared error (MSE) and accuracy. Because medium is closer to the midpoint on the intensity scale that we are using, the medium-class baseline performs better for MSE. The neutral-class baseline, on the other hand, performs better for accuracy, except for at the sentence level.

In Table 5.4, results for the same five experiments are given for each of the three classification algorithms. The experiments differ in which features and feature organizations are used. Experiment (1) in the table uses bag-of-words (BAG), where the words in each sentence are given to the classification algorithm as features. Experiments (2) and (3) use all the subjectivity clues described in Section 5.3. For experiment (2), the TYPE organization is used; for experiment (3), the INTENSITY organization is used. For experiments (4) and (5), bag-of-words is used along with the subjectivity clues in their two different feature organizations. The results in bold are the best for a particular clause level, experiment, and algorithm.

The results for intensity classification are promising for clauses at all levels of nesting. For BoosTexter, all experiments result in significant improvements over the two baselines, as measured by both MSE and accuracy. The same is true for Ripper, with the exception of experiment (1), which uses only bag-of-words and none of the subjectivity clue features. For SVM-light, at the sentence level (clause level 1), all experiments also result in significant

Table 5.4: Intensity classification results

	level 1		level 2		level 3		level 4	
Baselines	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc
neutral-class	3.603	28.1	2.752	41.8	2.539	45.9	2.507	48.3
medium-class	1.540	30.4	2.000	25.4	2.141	23.7	2.225	22.5

	level 1		level 2		level 3		level 4	
BoosTexter	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc
(1) BAG	1.234	50.9	1.390	53.1	1.534	53.6	1.613	53.0
(2) TYPE	1.135	50.2	1.267	53.4	1.339	54.7	1.410	55.5
(3) INTENSITY	1.060	54.1	1.180	56.9	1.258	**57.9**	1.269	**60.3**
(4) BAG + TYPE	1.069	52.0	1.178	54.8	1.267	55.9	1.321	56.8
(5) BAG + INTENSITY	**0.991**	**55.0**	**1.111**	**57.0**	**1.225**	57.5	**1.211**	59.4

	level 1		level 2		level 3		level 4	
Ripper	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc
(1) BAG	1.570	34.5	1.961	29.2	2.091	27.1	2.176	25.7
(2) TYPE	1.025	49.7	1.150	53.5	1.206	55.0	1.269	56.3
(3) INTENSITY	**0.999**	**53.2**	**1.121**	**55.6**	**1.181**	**56.1**	**1.205**	57.7
(4) BAG + TYPE	1.072	49.4	1.194	53.4	1.244	55.3	1.319	55.9
(5) BAG + INTENSITY	1.004	**53.2**	1.138	55.3	1.220	55.9	1.244	**57.8**

	level 1		level 2		level 3		level 4	
SVM-light	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc
(1) BAG	0.962	40.2	1.432	29.2	1.647	26.2	1.748	24.5
(2) TYPE	0.971	36.5	1.080	27.7	1.117	25.0	1.138	22.4
(3) INTENSITY	1.092	38.1	1.214	29.0	1.264	26.2	1.267	24.7
(4) BAG + TYPE	**0.750**	46.0	**0.926**	34.1	**1.023**	28.9	**1.065**	25.9
(5) BAG + INTENSITY	0.793	**48.3**	0.979	**36.3**	1.071	**32.1**	1.084	**29.4**

improvements over the baselines for MSE and accuracy. For the nested clause levels, all MSE results are significantly better than the MSE results provided by the more challenging medium-class baseline classifier. The same is not true, however, for the accuracy results, which are well below the accuracy results of the neutral-class baseline classifier.

The best experiments for all classifiers use all the subjectivity clues, supporting our hypothesis that using a wide variety of clues is effective. The experiment giving the best results varies somewhat for each classifier, depending on feature organization and whether BAG features are included. For BoosTexter, experiment (5) using BAG and INTENSITY features performs the best. For Ripper, experiment (3) using just the INTENSITY features performs the best, although not significantly better than experiment (5). For SVM-light, which experiment produces the best results depends on whether MSE or accuracy is the metric for evaluation. Experiment (4) using BAG and TYPE features has the better MSE results, experiment (5) using BAG and INTENSITY features has the better accuracies; the differences between the two experiments are significant (except for level-4 MSE).

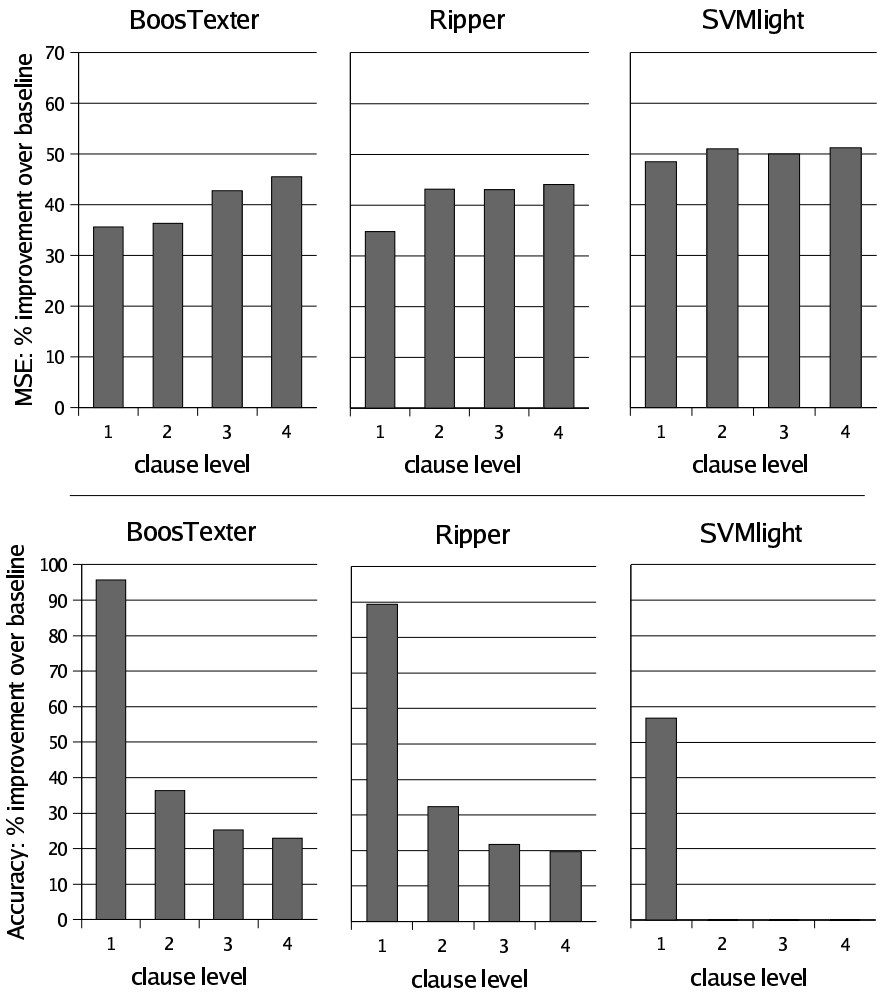
Figure 5.4 shows the percent improvements over baseline achieved by each classification algorithm for experiment (5). The medium-class baseline is used for MSE, and the neutral-class baseline is used for accuracy. For BoosTexter, the improvements in MSE range from 36% to 46%, and the improvements in accuracy range from 23% to 96%. The improvements over baseline for Ripper are similar. For SVM-light, the improvements over baseline for MSE are even better, close to 50% for all clause levels.

Note that BoosTexter and Ripper are non-ordinal classification algorithms, whereas support vector regression takes into account ordinal values. This difference is reflected in the results. The results are comparable for BoosTexter and Ripper (MSE is not significantly different; BoosTexter has slightly better accuracy). Although accuracies are lower, the regression algorithm achieves much better results for MSE. For experiment (5) using the BAG and INTENSITY features, SVM-light improves 10% to 20% over the MSE results for BoosTexter and 49% to 51% over baseline, coming closer to the true intensity at all clause levels.

5.5.3.1 Comparison with Upper Bound

In addition to baseline comparisons, it can also be informative to compare results to an upper bound. For intensity classification, the

Figure 5.4: Percent improvements over baseline for each algorithm for experiment (5)



upper bound on expected performance is the agreement level achieved by the annotators for marking intensity. Section 3.2.3.7 in Chapter 3 reported inter-annotator agreement for the intensity of sentences (level 1 in the experiments in this chapter). Average pairwise α-agreement was 0.77 with a 62% overall agreement. Average mean-squared error can also be calculated for the annotators' agreement: 0.617. These numbers give an idea for what the best performance is that can be expected for sentence-level intensity classification, given the current data.

Comparing the sentence-level MSE and accuracy results in Table 5.4 with the above levels of inter-annotator agreement shows that there is still quite a bit of room for improvement. Accuracy for level 1 ranges from 48.3% for SVM-light to 55% for BoosTexter. The best MSE achieved is 0.750 for SVM-light experiment (4). It is also possible to calculate Krippendorff's α for the classification experiments. For the best experiments for BoosTexter and Ripper, α is 0.56 and 0.55, respectively. The best SVM-light experiment has an α of 0.59.

5.5.3.2 Contribution of SYNTAX Clues In this section, I examine the contribution of the new syntax clues to the classification results. Table 5.5 shows the increases in MSE and the decreases in accuracy that result when the SYNTAX clues are omitted for experiment (5) (bag-of-words and INTENSITY feature organization).

Table 5.5 shows that the new SYNTAX clues do contribute information over and above bag-of-words and the clues from previous work (PREV clues). For all learning algorithms and all clause levels, omitting the SYNTAX clues results in a significant difference in MSE. The differences in accuracy are also significant, with the exception of BoosTexter levels 1 and 2 and Ripper level 4. The loss in accuracy for SVM-light, which already has lower accuracies, is particularly severe.

5.5.3.3 TYPE versus INTENSITY Feature Organization To examine the difference between the TYPE and INTENSITY feature organizations, I again turn to Table 5.4. For boosting, the experiments using the INTENSITY organization perform better, achieving lower mean-squared errors and higher accuracies. Comparing experiments (2) and (3), the INTENSITY organization performs significantly better at all clause levels. For

Table 5.5: Increases in MSE and decreases in accuracy that result when SYNTAX clues are omitted for experiment (5)

	Increase in MSE			
	level 1	level 2	level 3	level 4
BoosTexter	0.090	0.094	0.139	0.152
Ripper	0.226	0.209	0.238	0.215
SVM-light	0.056	0.185	0.229	0.262

	Decrease in Accuracy			
	level 1	level 2	level 3	level 4
BoosTexter	-0.9	-1.0	-2.1	-2.4
Ripper	-2.5	-1.8	-1.6	-1.2
SVM-light	-4.8	-5.1	-4.7	-4.2

experiments (4) and (5), improvements are again significant, with the exception of MSE for levels 3 and 4. For Ripper, experiments using the INTENSITY organization also achieve better results, although fewer improvements are significant. For SVM-light, the benefits of the INTENSITY organization are not as clear cut. Experiments using the INTENSITY organization all have higher accuracies, but their MSE is also worse. Furthermore, the differences are all significant, with the exception of the improvement in accuracy for experiment (3) level 3, and the increase in MSE for experiment (5) level 4. This makes it difficult to determine whether the INTENSITY organization is beneficial when performing support vector regression. For Ripper and BoosTexter, however, there is a clear benefit to using the INTENSITY organization for intensity classification.

5.6 RELATED WORK

To the best of my knowledge, this research is the first to automatically distinguish between not only subjective and objective (*neutral*) language, but among weak, medium, and strong subjectivity as well. The research most closely related is work by Yu and Hatzivassiloglou

(2003) and earlier work in this line of research (Wiebe, Bruce, and O'Hara, 1999; Riloff, Wiebe, and Wilson, 2003; Riloff and Wiebe, 2003) on classifying subjective and objective sentences. Yu and Hatzivassiloglou use Naïve Bayes classifiers to classify sentences as subjective or objective. The features they use include the words in each sentence, essentially bag-of-words, bigrams, trigrams, and counts of positive and negative words. Their sets of positive and negative words were learned starting with positive and negative adjectives from (Hatzivassiloglou and McKeown, 1997), which are included in the PREV clues used in the experiments in this chapter. Yu and Hatzivassiloglou also use clues that incorporate syntactic information, specifically clues that, for each sentence, encode the polarity of the head verb, main subject, and their modifiers. Unlike the syntactic clues used in this research, the syntactic clues from (Hatzivassiloglou and McKeown, 1997) did not help with their classifier's performance. Many of the PREV clues were originally used to classify subjective and objective sentences in (Riloff, Wiebe, and Wilson, 2003; Riloff and Wiebe, 2003). Wiebe et al. (1999) and Riloff et al. (2003) also used features that I did not use, including binary features to represent the presence of a pronoun, an adjective, a cardinal number, a modal other than *will*, and an adverb other than *not*.

Other researchers have worked to identify opinions below the sentence level (Kim and Hovy, 2004; Morinaga et al., 2002; Dave, Lawrence, and Pennock, 2003; Nasukawa and Yi, 2003; Yi et al., 2003; Hu and Liu, 2004; Popescu and Etzioni, 2005). Kim and Hovy (2004) identify sentences that mention particular topics, use a named entity tagger to identify the closest entity in the text, and then use the topic and entity phrases to define regions that are used for classifying sentiments. Dave *et al.* (2003), Nasukawa and Yi (2003), Yi *et al.* (2003), Hu and Liu (2004), and Popescu and Etzioni (2005) work on mining product reviews. In product review mining, the typical approach is to first identify references to particular products or product features of interest. Once these are identified, positive and negative opinions about the product are extracted. In contrast to the research above, the work in this chapter seeks to classify the intensity of nested clauses in all sentences in the corpus.

5.7 CONCLUSIONS

In this chapter, I performed experiments in automatically recognizing the intensity of sentences and clauses. Although the results for intensity classification are not high, they are promising, providing support for the hypotheses I am exploring in this dissertation. Classifiers trained using different learning algorithms achieved significant improvements over the baselines, demonstrating that automatic systems can be developed for performing fine-grained subjectivity analysis—fine grained both in terms of intensity and in terms of analysis below the sentence level.

I employed a wide range of features in these experiments, and the best performing classifiers for both boosting and SVM were those that used all the different features, both lexical and syntactic. Without the new syntactic features, performance drops, with many of the decreases being statistically significant. This provides evidence of the need for a wide variety of features for fine-grained subjectivity analysis.

I also hypothesized that a feature organization based on the intensity of clues would be beneficial for intensity classification. The experiments in this chapter provide some support for this hypothesis. Experiments using the INTENSITY feature organization performed the best for both BoosTexter and Ripper. Results for SVM were mixed, however, with experiments using the INTENSITY organization achieving higher accuracies, but not higher mean-squared errors.

6.0 RECOGNIZING CONTEXTUAL POLARITY

Sentiment analysis is a type of subjectivity analysis that focuses specifically on identifying positive and negative opinions, emotions, and evaluations. Although a great deal of the work in sentiment analysis has targeted documents, applications such as opinion question answering and review mining require a finer-grained level of analysis. For example, they must be able not only to pinpoint expressions of positive and negative sentiments, such as those underlined in sentence 6.1, but also to determine when an opinion is *not* being expressed by a word or phrase that typically does evoke one, such as "condemned" in sentence 6.2.

(6.1) African observers generally approved (**positive**) of his victory while Western governments denounced (**negative**) it.

(6.2) Gavin Elementary School was condemned in April 2004.

A common approach to sentiment analysis is to use a lexicon with information about which words and phrases are positive and which are negative. This lexicon may be manually compiled, as is the case with the General Inquirer (Stone et al., 1966), a resource often used in sentiment analysis. Alternatively, the information in the lexicon may be acquired automatically. Acquiring the polarity of words and phrases is itself an active line of research in the sentiment community, pioneered by the work of Hatzivassiloglou and McKeown (1997) on predicting the polarity or semantic orientation of adjectives. Various techniques have been proposed for learning the polarity of words. They include corpus-based techniques, such as using constraints on the co-occurrence of words with similar or opposite polarity in conjunctions (Hatzivassiloglou and McKeown, 1997) and statistical measures of word association (Turney and Littman, 2003), as well as techniques that exploit information about lexical relationships (Kamps and Marx, 2002; Kim and Hovy, 2004) and glosses (Esuli and

Sebastiani, 2005; Andreevskaia and Bergler, 2006) in resources such as WordNet.

Although acquiring the polarity of words and phrases is undeniably important, what the polarity of a given word or phrase is when it is used in a particular context is another problem entirely. Consider, for example, the underlined positive and negative words in the following sentence.

> (6.3) Philip Clapp, president of the National Environment Trust, sums up well the general thrust of the reaction of environmental movements: "There is no reason at all to believe that the polluters are suddenly going to become reasonable."

The first underlined word is "trust." Although many senses of the word "trust" express a positive sentiment, in this case, the word is not being used to express a sentiment at all. It is simply part of an expression referring to an organization that has taken on the charge of caring for the environment. The adjective "well" is considered positive, and indeed it is positive in this context. However, the same is not true for the words "reason" and "reasonable." Out of context, both of these words are considered positive. In context, the word "reason" is being negated, changing its polarity from positive to negative. The phrase "no reason at all believe" changes the polarity of the proposition that follows; because "reasonable" falls within this proposition, its polarity becomes negative. In the context of this article, the word "polluters" is similar to the word "trust" in that it is mainly being used as a referring expression (to companies that pollute).

I use the term **prior polarity** to refer to the polarity that would be listed for a word in a lexicon, and the term **contextual polarity** to refer the polarity of the expression in which a word appears, considering the context of the sentence and document. Although words often do have the same prior and contextual polarity, many times the word's prior and contextual polarities differ. Words with a positive prior polarity may have a negative contextual polarity, or vice versa. Also, words that are positive or negative out of context quite often are *neutral* in context, meaning that they are not even being used to express a sentiment.

The focus of this chapter is the recognition of contextual polarity. I begin by describing an annotation scheme for marking sentiment expressions and their contextual polarity in the MPQA Corpus and an inter-annotator agreement study conducted by Paul Hoffmann.

The results of this study show that, given a set of subjective expressions identified from the existing annotations in the MPQA Corpus, contextual polarity can be reliably annotated.

Using the contextual polarity annotations, I conduct experiments in automatically distinguishing between prior and contextual polarity. Beginning with a large lexicon of clues tagged with their prior polarity, I identify the contextual polarity of the instances of those clues in the corpus. The process that I use has two steps, first classifying each clue as being in a neutral or polar phrase, and then disambiguating the contextual polarity of the clues marked as polar. For each step in the process, I experiment with a variety of features and evaluate the performance of the features using several different machine learning algorithms.

The experiments in this chapter reveal a number of interesting findings. First, being able to accurately identify neutral contextual polarity, when a sentiment clue is *not* being used to express a sentiment, is an important aspect to the problem. The importance of neutral examples has previously been noted for classifying the sentiment of documents (Koppel and Schler, 2006), but this is the first work to explore how neutral instances affect classifying the contextual polarity of words and phrases. In particular, I found that the performance of features for distinguishing between positive and negative polarity greatly degrades when neutral instances are included in the experiments.

I also found that achieving the best performance for recognizing contextual polarity requires a wide variety of features. This is particularly true for distinguishing between neutral and polar instances. Although some features help to increase polar or neutral recall or precision, it is only the combination of features together that achieve significant improvements in accuracy over the baselines. The experiments show that for distinguishing between positive and negative instances, features capturing negation are clearly the most important. However, there is more to the story than simple negation. Features that capture relationships between instances of clues also performed well, indicating that identifying features that represent more complex interdependencies between polarity clues may be an important avenue for future research.

The remainder of this chapter is organized as follows. Section 6.1 gives an overview of some of the things that can influence contextual polarity. The annotation scheme and inter-annotator agreement study for contextual polarity are described in Section 6.2. In Sections

6.3 and 6.4 I describe the lexicon used in the experiments in this chapter, and how the contextual polarity annotations are used to determine the gold standard tags for instances from the lexicon. In Section 6.5 I consider what kind of performance can be expected from a simple prior-polarity classifier. Section 6.6 describes the features that are used in the contextual polarity experiments, and the experiments are presented in Section 6.7. In Section 6.8 I briefly discuss related work, and in Section 6.9 I conclude.

6.1 POLARITY INFLUENCERS

Phrase-level sentiment analysis is not a simple problem. Many things besides negation can influence contextual polarity, and even negation is not always straightforward. Negation may be local (e.g., **not** *good*), or involve longer-distance dependencies such as the negation of the proposition (e.g., **does not** *look very good*) or the negation of the subject (e.g., **no one** *thinks that it's good*). In addition, certain phrases that contain negation words intensify rather than change polarity (e.g., **not only** *good but amazing*). Contextual polarity may also be influenced by modality (e.g., whether the proposition is asserted to be real (*realis*) or not real (*irrealis*) – *no reason at all to believe* is irrealis, for example); word sense (e.g., *Environmental* **Trust** versus *He has won the people's* **trust**); the syntactic role of a word in the sentence (e.g., whether the word is in the subject or objective of a copular verb, consider **polluters** *are* versus *they are* **polluters**); and diminishers such as *little* (e.g., **little** *truth*, **little** *threat*). Polanyi and Zaenen (2004) give a detailed discussion of many of the above types of polarity influencers. Many of these contextual polarity influencers are represented as features in the machine learning experiments in this chapter.

Contextual polarity may also be influenced by things such as the domain or topic. For example, the word *cool* is positive if used to describe a car, but it is negative if it is used to describe someone's demeanor. Similarly, a word such as *fever* is unlikely to be expressing a sentiment when used in a medical context. One feature is used in the experiments to represent the topic of the document.

Another important aspect of contextual polarity is the perspective of the person who

is expressing the sentiment. For example, consider the phrase "failed to defeat" in the sentence *Israel failed to defeat Hezbollah*. From the perspective of Israel, "failed to defeat" is negative. From the perspective of Hezbollah, "failed to defeat" is positive. Therefore, the contextual polarity of this phrase ultimately depends on the perspective of who is expressing the sentiment. Although automatically detecting this kind of pragmatic influence on polarity is beyond the scope of this research, this as well as the other types of polarity influencers all were considered when annotating contextual polarity.

6.2 CONTEXTUAL POLARITY ANNOTATIONS

Investigating the contextual polarity of sentiment expressions in the MPQA Corpus requires new annotations. Although the polarity of a subset of expressions was captured with the *attitude-type* attribute, this attribute was not comprehensively annotated (see Chapter 3). However, **subjective expressions** in the corpus were comprehensively annotated. Subjective expressions in the MPQA Corpus are a subset of the private state annotations. They include all expressive subjective element frames and those direct subjective frames with an *expression intensity* greater than neutral. Because sentiment is a type of private state, sentiment expressions will be a subset of the subjective expressions already marked in the corpus. Thus, the subjective expression annotations in the MPQA Corpus give a starting point for the sentiment and contextual polarity annotations.

6.2.1 Annotation Scheme

When deciding how to annotate contextual polarity, there where two main issues that needed to be addressed. First, which of the subjective expressions are sentiment expressions? Second, what annotation scheme should be used for marking contextual polarity?

For this research, sentiments are defined as positive and negative emotions, evaluations and stances. Examples of positive sentiments are on the left in Table 6.1, and examples of negative sentiments are on the right. Any subjective expression that is expressing one of

Table 6.1: Examples of positive and negative sentiments

	Positive sentiments	Negative sentiments
Emotion	I'm happy	I'm sad
Evaluation	Great idea!	Bad idea!
Stance	She supports the bill	She's against the bill

these types of private states is considered a sentiment expression.

The second issue to address is what the actual annotation scheme should be for marking contextual polarity. The scheme that was developed has four tags: *positive*, *negative*, *both*, and *neutral*. The *positive* tag is used to mark positive sentiments. The *negative* tag is used to mark negative sentiments. The *both* tag is applied to sentiment expressions where both a positive and negative sentiment are being expressed (e.g., a *bittersweet* memory). The *neutral* tag is used for all other subjective expressions.

Below are examples of contextual polarity annotations from the corpus. The tags are in boldface, and the subjective expressions with the given tags are underlined.

(6.4) Thousands of coup supporters celebrated (**positive**) overnight, waving flags, blowing whistles ...

(6.5) The criteria set by Rice are the following: the three countries in question are repressive (**negative**) and grave human rights violators (**negative**) ...

(6.6) Besides, politicians refer to good and evil (**both**) only for purposes of intimidation and exaggeration.

(6.7) Jerome says the hospital feels (**neutral**) no different than a hospital in the states.

As a final note on the annotation scheme, the annotators were asked to judge the contextual polarity of the sentiment that was ultimately being conveyed by the subjective expression, that is, once the sentence had been fully interpreted. Thus, the subjective expression, "they have not succeeded, and will never succeed," was marked as positive in the following sentence:

(6.8) They have not succeeded, and will never succeed (**positive**), in breaking the will of this valiant people.

Table 6.2: Contingency table for contextual polarity agreement

	Neutral	Positive	Negative	Both	Total
Neutral	**123**	14	24	0	161
Positive	16	**73**	5	2	96
Negative	14	2	**167**	1	184
Both	0	3	0	**3**	6
Total	153	92	196	6	**447**

The reasoning is that breaking the will of a valiant people is negative, so to not succeed in breaking their will is positive.

6.2.2 Agreement Study

Paul Hoffmann conducted an agreement study to measure the reliability of the polarity annotation scheme. For the study, two annotators[1] independently annotated 10 documents from the MPQA Corpus containing 447 subjective expressions. Table 6.2 shows the contingency table for the two annotators' judgments. Overall agreement is 82%, with a Kappa (κ) value of 0.72.

As part of the annotation scheme, annotators were asked to judge how certain they were in their polarity tags. For 18% of the subjective expressions, at least one annotator used the *uncertain* tag when marking polarity. If these cases are considered borderline and excluded from the study, percent agreement increases to 90% and Kappa rises to 0.84. Table 6.3 shows the revised contingency table with the uncertain cases removed. This shows that annotator agreement is especially high when both annotators are certain, and that annotators are certain for over 80% of their tags.

Note that all annotations are included in the experiments.

[1] Paul Hoffmann and myself.

Table 6.3: Contingency table for contextual polarity agreement with borderline cases removed

	Neutral	Positive	Negative	Both	Total
Neutral	**113**	7	8	0	128
Positive	9	**59**	3	0	71
Negative	5	2	**156**	1	164
Both	0	2	0	**2**	4
Total	127	70	167	3	**367**

Table 6.4: Distribution of contextual polarity tags

Neutral	Positive	Negative	Both	Total
9,057	3,311	7,294	299	19,961
45.4%	16.6%	36.5%	1.5%	100%

6.2.3 MPQA Corpus version 1.2

In total, all 19,962 subjective expressions in the 535 documents (11,112 sentences) of the MPQA Corpus were annotated with their contextual polarity as described above. Table 6.4 gives the distribution of the tags. This table shows that a small majority of subjective expressions (54.6%) are expressing a positive, negative, or both (positive and negative) sentiment. I refer to these expressions as **polar in context**. Close to half of the subjective expressions are neutral: They are expressing some other type of subjectivity other than sentiment. This suggests that, although sentiment is a major type of subjectivity, there are other prominent types of subjectivity that may be important to distinguish for applications seeking to exploit subjectivity analysis.

As many NLP applications operate at the sentence level, one important issue to consider is the distribution of sentences with respect to the subjective expressions they contain. In the 11,112 sentences in the MPQA corpus, 28% contain no subjective expressions, 24% contain only one, and 48% contain two or more. Of the 5,304 sentences containing two or more subjective expressions, 17% contain mixtures of positive and negative expressions, and 61%

contain mixtures of polar (positive/negative/both) and neutral subjective expressions.

6.3 PRIOR-POLARITY SUBJECTIVITY LEXICON

The lexicon that I use for the experiments in this chapter is a collection of over 8,000 single-word subjectivity clues. The majority of the clues come from the lists of subjectivity clues used in (Riloff, Wiebe, and Wilson, 2003).[2] In (Riloff and Wiebe, 2003), these words are grouped according to their reliability as subjectivity clues. Words that are subjective in most contexts were marked strongly subjective (*strongsubj*), and those that may only have certain subjective usages were marked weakly subjective (*weaksubj*). This reliability class information is retained in the lexicon.

With the help of Paul Hoffmann, the lexicon was expanded and the clues were tagged with their prior polarity. To expand the lexicon, additional potentially subjective words were identified from the General Inquirer positive and negative word lists (Stone et al., 1966) and with the help of a dictionary and thesaurus. The newly added words were also given reliability tags, either *strongsubj* or *weaksubj*. The final lexicon has a coverage of 67% of the subjective expressions in the MPQA Corpus, where coverage is the percentage of subjective expressions containing one or more instances of clues from the lexicon.

The next step was to tag the clues in the lexicon with their prior polarity. For words that came from positive and negative word lists (Stone et al., 1966; Hatzivassiloglou and McKeown, 1997), their original polarity, either *positive* or *negative*, was largely retained. The remaining words were assigned one of the tags *positive*, *negative*, *both* or *neutral*.

By far, the majority of clues, 92.8%, are marked as having either positive (33.1%) or negative (59.7%) prior polarity. Only a small number of clues (0.3%) are marked as having both positive and negative polarity. These are words like *brag*, where the one who is bragging is expressing something positive, but describing someone as bragging is expressing a negative evaluation of that person. 6.9% of the clues in the lexicon are marked as neutral. Examples of these are verbs such as *feel*, *look*, and *think*, and intensifiers such as *deeply*, *entirely*, and

[2]These clues are a subset of the clues used in Chapter 5.

practically. These words are included because, although their prior polarity is neutral, they are good clues that a sentiment is being expressed (e.g., **feels** *slighted*, **feels** *satisfied*, **look** *kindly on*, **look** *forward to*). Including them increases the coverage of the system.

At the end of the previous section, I considered the distribution of sentences in the MPQA Corpus with respect to the subjective expressions they contain. It's interesting to compare that distribution with the distribution of sentences with respect to the clues they contain from the lexicon. More sentences have two or more clue instances (62%) than have two or more subjective expressions (48%). More importantly, many more sentences have mixtures of positive and negative clues than actually have mixtures of positive and negative subjective expressions. Only 880 sentences have a mixture of both positive and negative subjective expressions, while 3,234 sentences have mixtures of positive and negative clues. This strongly suggests that being able to disambiguate the contextual polarity of subjectivity and sentiment clues is an important aspect of higher-level tasks, such as classifying the sentiment of sentences.

6.4 DEFINITION OF THE GOLD STANDARD

In the experiments described in the following sections, the goal is to classify the contextual polarity of the expressions that contain instances of the subjectivity clues in the lexicon. However, determining which clue instances are part of the same expression and identifying expression boundaries are not the focus of this work. Thus, instead of trying to identify and label each expression, in the experiments below, each clue instance is labelled individually as to its contextual polarity.

I define the gold-standard contextual polarity of a clue instance in terms of the manual annotations (Section 6.2) as follows. If a clue instance is not in a subjective expression (and therefore not in a sentiment expression), its gold class is *neutral*. If a clue instance appears in just one subjective expression or in multiple subjective expressions with the same contextual polarity, its gold class is the contextual polarity of the subjective expression(s). If a clue appears in a mixture of negative and neutral subjective expressions, its gold class is *negative*;

Table 6.5: Confusion matrix for the prior-polarity classifier on the development set

		Prior-Polarity Classifier				
		Neutral	Positive	Negative	Both	Total
Gold Class	Neutral	**798**	784	698	4	2284
	Positive	81	**371**	40	0	492
	Negative	149	181	**622**	0	952
	Both	4	11	13	**5**	33
	Total	1032	1347	1373	9	**3761**

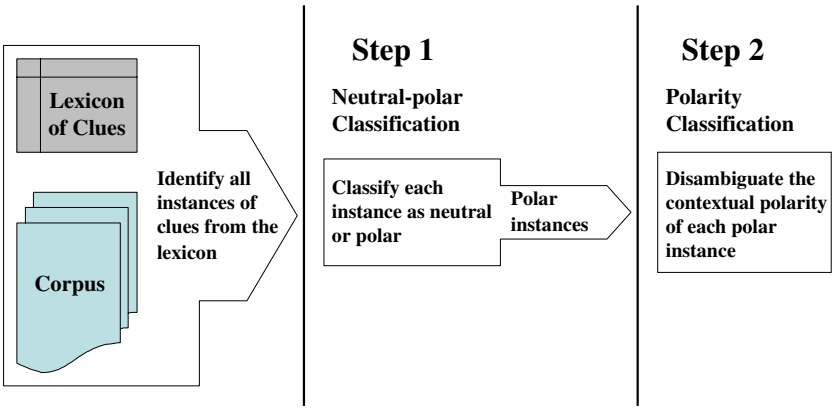
if it is in a mixture of positive and neutral subjective expressions, its gold class is *positive*. Finally, if a clue appears in at least one positive and one negative subjective expression (or in a subjective expression marked as *both*), then its gold class is *both*.

6.5 A PRIOR-POLARITY CLASSIFIER

Before delving into the task of recognizing contextual polarity, an important question to address is how useful prior polarity alone is for identifying contextual polarity. To answer this question, I created a classifier that simply assumes the contextual polarity of a clue instance is the same as the clue's prior polarity. I explored this classifier's performance on a small amount of development data, which is not part of the data used in the experiments below.

This simple classifier has an accuracy of 48%. The confusion matrix given in Table 6.5 shows that 76% of the errors result from words with non-neutral prior polarity appearing in phrases with neutral contextual polarity. Only 12% of the errors result from words with neutral prior polarity appearing in expressions with non-neutral contextual polarity, and only 11% of the errors come from words with a positive or negative prior polarity appearing in expressions with the opposite contextual polarity. Table 6.5 also shows that positive clues tend to be used in negative expressions far more often than negative clues tend to be used in positive expressions.

Figure 6.1: Two-step approach to recognizing contextual polarity



Given that by far the largest number of errors come from clues with *positive*, *negative*, or *both* prior polarity appearing in *neutral* contexts, I was motivated to try a two-step approach to the problem of sentiment classification. The first step, **Neutral-Polar Classification**, tries to determine if an instance is neutral or polar in context. The second step, **Polarity Classification**, takes all instances that step one classified as polar, and tries to disambiguate their contextual polarity. This two-step approach is illustrated in Figure 6.1.

6.6 FEATURES

The features I use in the contextual polarity experiments were motivated both by the literature and by exploration of the contextual polarity annotations in the development data. A number of features were inspired by the paper by Polanyi and Zaenen (2004) on contextual polarity influencers. Other features are those that have been found useful in the past for recognizing subjective sentences (Riloff, Wiebe, and Wilson, 2003; Wiebe, Bruce, and O'Hara, 1999).

6.6.1 Features for Neutral-Polar Classification

For distinguishing between neutral and polar instances, I use the features listed in Table 6.6. For ease of description, I group the features into 6 sets: word features, general modification features, polarity modification features, structure features, sentence features, and one document feature.

Word Features: In addition to the word token (the token of the clue instance from the lexicon) the word features include parts-of-speech of the previous word, the word itself, and the next word. The *prior polarity* and *reliability class* features represent those pieces of information about the clue, which are taken from the lexicon.

General Modification Features: These are binary features that capture different types of relationships involving the clue instance.

The first four features involve relationships with the word immediately before or after the clue. The *preceded by adjective* feature is true if the clue is a noun preceded by an adjective. The *preceded by adverb* feature is true if the preceding word is an adverb other than *not*. The *preceded by intensifier* feature is true if the preceding word is an intensifier, and the *self intensifier* feature is true if the clue itself is an intensifier. A word is considered to be an intensifier if it appears in a list of intensifiers and if it precedes a word of the appropriate part-of-speech (e.g., an intensifier adjective must come before a noun).

The *modify* features involve the dependency parse tree of the sentence. Parse trees are obtained as described in Section 5.3.2 of the previous chapter. Figure 6.2 gives the dependency parse tree for the sentence: *The human rights report poses a substantial challenge to the US interpretation of good and evil.* Each instance of a subjectivity clue from the lexicon is marked with the clue's prior polarity and reliability class.

For each clue instance, the *modify* features capture whether there are *adj*, *mod*, or *vmod* relationships between the clue instance and any other instances from the lexicon. Specifically, the *modifies strongsubj* feature is true if the clue and its parent share an *adj*, *mod*, or *vmod* relationship, and if its parent is a clue from the lexicon with *strongsubj* reliability. The *modifies weaksubj* feature is the same, except that it looks for clues with *weaksubj* reliability in the parent. The *modified by strongsubj* feature is true for a clue if one of its children is

Table 6.6: Features for neutral-polar classification

Word Features
word token
word part-of-speech
previous word part-of-speech
next word part-of-speech
prior polarity: positive, negative, both, neutral
reliability class: strongsubj or weaksubj
General Modification Features
preceded by adjective: binary
preceded by adverb (other than not): binary
preceded by intensifier: binary
self intensifier: binary
modifies strongsubj: binary
modifies weaksubj: binary
modified by strongsubj: binary
modified by weaksubj: binary
Polarity Modification Features
modifies polarity: positive, negative, neutral, both, notmod
modified by polarity: positive, negative, neutral, both, notmod
conjunction polarity: positive, negative, neutral, both, notmod
Structure Features
in subject: binary
in copular: binary
in passive: binary
Sentence Features
strongsubj clues in current sentence: 0, 1, 2, 3 (or more)
strongsubj clues in previous sentence: 0, 1, 2, 3 (or more)
strongsubj clues in next sentence: 0, 1, 2, 3 (or more)
weaksubj clues in current sentence: 0, 1, 2, 3 (or more)
weaksubj clues in previous sentence: 0, 1, 2, 3 (or more)
weaksubj clues in next sentence: 0, 1, 2, 3 (or more)
adjectives in sentence: 0, 1, 2, 3 (or more)
adverbs in sentence (other than not): 0, 1, 2, 3 (or more)
cardinal number in sentence: binary
pronoun in sentence: binary
modal in sentence (other than will): binary
Document Feature
document topic

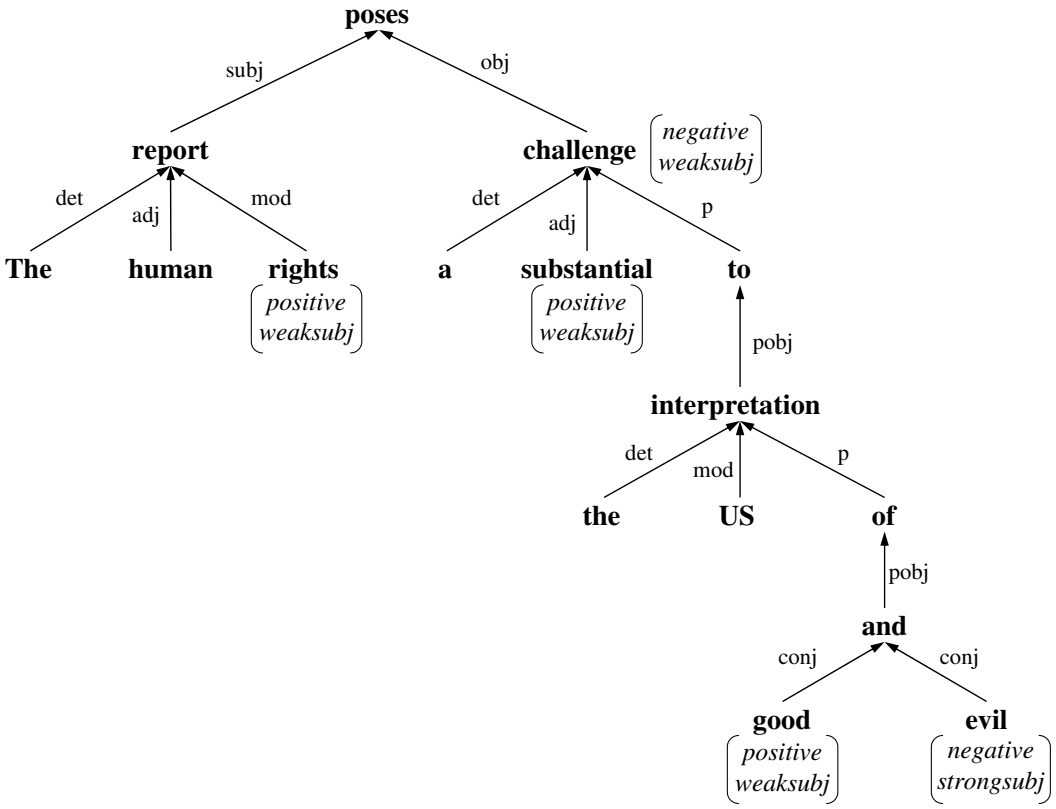
a clue with *strongsubj* reliability, and if the clue and its child share an *adj*, *mod*, or *vmod* relationship. The *modified by weaksubj* feature is the same, except that it looks for *weaksubj* clues in the children. Although the *adj* and *vmod* relationships are typically local, the *mod* relationship involves longer-distance as well as local dependencies. Figure 6.2 helps to illustrate these features. The *modifies weaksubj* feature is true for the clue "substantial," but false for the clue "rights." The *modified by weaksubj* feature is false for the clue "substantial," but true for the clue "challenge."

Polarity Modification Features: The *modifies polarity*, *modified by polarity*, and *conj polarity* features capture specific relationships between the clue instance and other polarity clues. If the clue and its parent in the dependency tree share an *obj*, *adj*, *mod*, or *vmod* relationship, the *modifies polarity* feature is set to the prior polarity of the clue's parent. If the parent is not in the prior-polarity lexicon, its prior polarity is considered *neutral*. If the clue is at the root of the tree and has no parent, the value of the feature is *notmod*. The *modified by polarity* feature is similar, looking for *adj*, *mod*, and *vmod* relationships and polarity words within the clue's children. The *conj polarity* feature determines if the clue is in a conjunction. If so, the value of this feature is its sibling's prior polarity. As above, if the sibling is not in the lexicon, its prior polarity is neutral. If the clue in not in a conjunction, the value for this features is *notmod*. Figure 6.2 also helps to illustrate these modification features: *modifies polarity* is negative for the word "substantial," *modified by polarity* is positive for the word "challenge," and *conj polarity* is negative for the word "good" and positive for the word "evil."

Structure Features: These are binary features that are determined by starting with the clue instance and climbing up the dependency parse tree toward the root, looking for particular relationships, words, or patterns. The *in subject* feature is true if there is a *subj* relationship on the path to the root. The *in copular* feature is true if *in subject* is false and if a node along the path is both a main verb and a copular verb. The *in passive* feature is true if a passive verb pattern is found on the climb.

The *in subject* and *in copular* features were motivated by the intuition that the syntactic role of a word may influence whether a word is being used to express a sentiment. For example, consider the word "polluters" in each of the following two sentences.

Figure 6.2: Dependency parse tree for the sentence, *The human rights report poses a substantial challenge to the US interpretation of good and evil*, with prior polarity and reliability class indicated for instances of clues from the lexicon



(6.9) Under the application shield, **polluters** are allowed to operate if they have a permit.

(6.10) "The big-city folks are pointing at the farmers and saying you are **polluters** ..."

In the first sentence, "polluters" is simply being used as a referring expression. In the second sentence, "polluters" clearly is being used to express a negative evaluation of the farmers.

The motivation for the *in passive* feature was previous work by Riloff and Wiebe (2003), who found that different words are more or less likely to be subjective depending on whether they are in the active or passive.

Sentence Features: These are features that previously were found useful for sentence-level subjectivity classification (Riloff, Wiebe, and Wilson, 2003; Wiebe, Bruce, and O'Hara, 1999). They include counts of strongsubj and weaksubj clues in the current, previous and next sentences, counts of adjectives and adverbs other than *not* in the current sentence, and binary features to indicate whether the sentence contains a pronoun, a cardinal number, and a modal other than *will*.

Document Feature: There is one document feature representing the topic of the document. The motivation for this feature is that whether or not a word is expressing a sentiment or is even subjective may depend on the subject of the discourse. For example, the words "fever" and "sufferer" may express a negative sentiment in certain contexts, but probably not in a heath or medical context, as is the case in the following sentence.

(6.11) The disease can be contracted if a person is bitten by a certain tick or if a person comes into contact with the blood of a congo **fever sufferer**.

About two thirds of the documents in the MPQA Corpus were already labelled with one of the 10 topics previously listed in Table 2.1. The remaining documents were labelled with one of the following five general topics: economics, general politics, health, report events, and war and terrorism.

6.6.2 Features for Polarity Classification

Table 6.6.2 lists the features that I use for step 2, polarity classification. *Word token*, *word prior polarity*, and the *polarity modification* features are the same as described for neutral-

Table 6.7: Features for polarity classification

Word Features
word token
word prior polarity: positive, negative, both, neutral
Negation Features
negated: binary
negated subject: binary
Polarity Modification Features
modifies polarity: positive, negative, neutral, both, notmod
modified by polarity: positive, negative, neutral, both, notmod
conj polarity: positive, negative, neutral, both, notmod
Polarity Shifters
general polarity shifter: binary
negative polarity shifter: binary
positive polarity shifter: binary

polar classification.

I use two features to capture two different types of negation. The *negated* feature is a binary feature that captures more local negations: Its value is true if a negation word or phrase is found within the four preceding words, and if the negation word is not also in a phrase that acts as an intensifier rather than a negator. Examples of phrases that intensify rather than negate are *not only* and *nothing if not*. The *negated subject* feature captures a longer-distance type of negation. This features is true if the subject of the clause containing the word is negated. For example, the *negated subject* feature is true for the word "support" in the following sentence.

(6.12) No politically prudent Israeli could **support** either of them.

The last three polarity features look in a window of four words before, searching for the presence of particular types of polarity influencers. *General polarity shifters* reverse polarity (e.g., *little* truth, *little* threat). *Negative polarity shifters* typically make the polarity of an

expression negative (e.g., *lack* of understanding). *Positive polarity shifters* typically make the polarity of an expression positive (e.g., *abate* the damage).

6.7 EXPERIMENTS IN RECOGNIZING CONTEXTUAL POLARITY

I have two primary goals with the following experiments in recognizing contextual polarity. The first is to evaluate the features described in Section 6.6 as to their usefulness for this task. The second is to investigate the importance of recognizing neutral instances—recognizing when a clue is not being used to express a sentiment—for classifying contextual polarity.

To evaluate the features, I investigate their performance, both together and in separate sets, across all four of the algorithms described in Chapter 4: BoosTexter, TiMBL IB1, Ripper, and SVM-light/SVM-multiclass. SVM-light is used for the experiments involving binary classification (neutral-polar classification), and SVM-multiclass is used for experiments with more than two classes.

For all of the classification algorithms except for SVM, the features are represented as they are presented in Section 6.6. For SVM, the representations for numeric and discrete-valued features are changed. Numeric features, such as the count of *strongsubj* clue instances in a sentence, are scaled to range between 0 and 1. Discrete-valued features, such as the *reliability class* feature, are converted into multiple binary features. For example, the *reliability class* feature is represented by two binary features: one for whether the clue is strongsubj and one for whether the clue is weaksubj.

To investigate the importance of recognizing neutral instances, I perform two sets of polarity classification (step 2) experiments. First, I experiment with classifying the polarity of all gold-standard polar instances—the clue instances identified as polar in context by the manual polarity annotations. Second, I experiment with using the polar instances identified automatically by the neutral-polar classifiers. Because the second set of experiments includes the neutral instances misclassified in step 1, results for the two sets of experiments can be compared to see how the noise of neutral instances affects the performance of the various polarity features.

All experiments are performed using 10-fold cross-validation over a test set of 10,287 sentences from 494 MPQA Corpus documents. I measure performance in terms of accuracy, recall, precision, and F-measure. Recall, precision, and F-measure for a given class C are defined as follows. Recall is the percentage of all instances of class C correctly identified by the classifier.

$$Rec(C) = \frac{\mid \text{instances of C correctly identified} \mid}{\mid \text{all instances of C} \mid}$$

Precision is the percentage of instances identified as class C by the classifier that are class C in truth.

$$Prec(C) = \frac{\mid \text{instances of C correctly identified} \mid}{\mid \text{all instances identified as C} \mid}$$

F-measure is the harmonic mean of recall and precision.

$$F(C) = \frac{2 \times Rec(C) \times Prec(C)}{Rec(C) + Prec(C)}$$

All results reported are averages over the 10 folds.

6.7.1 Neutral-Polar Classification

In the two-step process for recognizing contextual polarity, the first step is neutral-polar classification, determining whether each instance of a clue from the lexicon is neutral or polar in context. In the test set, there are 26,729 instances of clues from the lexicon. The features that are used for this step were listed above in Table 6.6 and described in Section 6.6.1.

In this section, I perform two sets of experiments. In the first, I compare the results of neutral-polar classification using all the neutral-polar features against the baselines, and in the second set of experiments, I explore the performance of individual sets of features. For the baselines, I use a classifier trained using just the *word token* feature, as well as a classifier (word+priorpol) trained using the *word token* and *prior polarity* features. These are challenging, but very appropriate baselines. The word token and the prior polarity of the clue instance represent a considerable amount of starting knowledge for this task, and

Table 6.8: Algorithm settings for neutral-polar classification

Algorithm	Settings
BoosTexter	2000 rounds of boosting
TiMBL	k=25, MVDM distance metric
Ripper	-!n, -S 0.5
SVM	linear kernel

both baselines do much better than choosing the most frequent class or classifying instances according to their prior polarity (Section 6.5). However, the information represented by these baselines is exactly what is need as a point for comparison to evaluate whether there is additional utility in the features proposed for neutral-polar classification.

To determine the parameter settings for the machine learning algorithms, I performed 10-fold cross-validation of the more challenging baseline classifier (word+priorpol) on the development data, varying select parameter settings. The results from those experiments were then used to select the parameter settings for the experiments on the test data. For BoosTexter, I varied the number of rounds of boosting. For TiMBL, I varied the value for k (the number of neighbors) and the distance metric (overlap or modified value difference metric (MVDM)[3]). For Ripper, I varied whether negative tests were disallowed for nominal (-!n) and set (-!s) valued attributes and how much to simplify the hypothesis (-S). For SVM, I experimented with linear, polynomial, and radial basis function kernels. Table 6.8 gives the settings selected for the neutral-polar classification experiments for the different learning algorithms.

6.7.1.1 Classification Results The results for the first set of experiments are given in Table 6.9. For each algorithm, I give the results for the two baseline classifiers, followed by

[3] Overlap is the most basic distance metric. The distance between two instances is simply the sum of the differences between the features. For features with symbolic values, the distance is 0 if there is an exact match between the values; otherwise, the distance is 1. Distance for the modified value difference metric (MVDM) is more complex. This metric measures the difference between two feature values by looking at their co-occurrence with the target classes. If the conditional distribution of the target classes given two feature values is similar, the distance between the two values will be low. More information about these two metrics can be found in the TiMBL Reference Guide (Daelemans et al., 2003b).

Table 6.9: Results for Step 1 Neutral-Polar Classification

		Polar			Neutral		
	Acc	Rec	Prec	F	Rec	Prec	F
BoosTexter							
word token baseline	74.0	41.9	77.0	54.3	92.7	73.3	81.8
word+priorpol baseline	75.0	55.6	70.2	62.1	86.2	76.9	81.3
neutral-polar features	**76.5**	**58.3**	72.4	**64.6**	87.1	**78.2**	82.4
TiMBL							
word token baseline	74.6	47.9	73.9	58.1	90.1	74.8	81.8
word+priorpol baseline	74.6	48.2	73.7	58.3	90.0	74.9	81.7
neutral-polar features	**76.5**	**59.5**	71.7	**65.0**	86.3	**78.5**	82.3
Ripper							
word token baseline	66.3	11.2	80.6	19.6	98.4	65.6	78.7
word+priorpol baseline	65.5	07.7	84.5	14.1	99.1	64.8	78.4
neutral-polar features	**71.4**	**49.4**	64.6	**56.0**	84.2	**74.1**	78.8
SVM							
word token baseline	74.6	47.9	73.9	58.1	90.1	74.8	81.8
word+priorpol baseline	75.6	54.5	72.5	62.2	88.0	76.8	82.0
neutral-polar features	75.3	52.6	72.7	61.0	88.5	76.2	81.9

the results for the classifier trained using all the neutral-polar features. The results shown in bold are significantly better than both baselines (two-sided t-test, $p < 0.05$) for the given algorithm. For SVM, the best classifier is the word+priorpol baseline.

Working together, how well do the neutral-polar features perform? For BoosTexter, TiMBL, and Ripper, the classifiers trained using all the features improve significantly over the two baselines in terms of accuracy, polar recall, polar F-measure, and neutral precision. Neutral F-measure is also higher, but not significantly so. These consistent results across three of the four algorithms show that the neutral-polar features are helpful for determining when a sentiment clue is actually being used to express a sentiment.

Interestingly, Ripper is the only algorithm for which the word-token baseline performed better than the word+priorpol baseline. Nevertheless, the *prior polarity* feature is an important component in the performance of the Ripper classifier using all the features. Excluding prior polarity from this classifier results in a significant decrease in performance for every metric. Decreases range from from 2.5% for neutral recall to 9.5% for polar recall.

Table 6.10: Neutral-polar feature sets for evaluation

Experiment	Features
PARTS-OF-SPEECH	parts-of-speech for clue instance, previous word, and next word
RELIABILITY-CLASS	reliability class of clue instance
PRECEDED-POS	preceded by adjective, preceded by adverb
INTENSIFY	preceded by intensifier, self intensifier
RELCLASS-MOD	modifies strongsubj/weaksubj, modified by strongsubj/weaksubj
POLARITY-MOD	polarity modification features
STRUCTURE	structure features
CURSENT-COUNTS	strongsubj/weaksubj clue instances in sentence
PNSENT-COUNTS	strongsubj/weaksubj clue instances in previous/next sentence
CURSENT-OTHER	adjectives/adverbs/cardinal number/pronoun/modal in sentence
TOPIC	document topic

The best SVM classifier is the word+priorpol baseline. In terms of accuracy, this classifier does not perform much worse than the BoosTexter and TiMBL classifiers that use all the neutral-polar features: The SVM word+priorpol baseline classifier has an accuracy of 75.6%, and both the BoosTexter and TiMBL classifiers have an accuracy of 76.5%. However, the BoosTexter and TiMBL classifiers that use all the features perform notably better in terms of polar recall and F-measure. The BoosTexter and TiMBL classifiers have polar recalls that are 8.6% and 11% higher than SVM. Polar F-measures for BoosTexter and TiMBL are 6.3% and 7.4% higher than for SVM. These increases are significant for $p < 0.01$.

6.7.1.2 Feature Set Evaluation To evaluate the contribution of the various features for neutral-polar classification, I performed a series of experiments in which different sets of neutral-polar features are added to the word+priorpol baseline and new classifiers are trained. I then compared the performance of these new classifiers to the word+priorpol baseline, with the exception of the Ripper classifiers, which I compared to the higher word baseline. Table 6.10 lists the sets of features tested in these experiments. The features sets generally correspond how the neutral-polar features are presented in Table 6.6, although some of the groups are broken down into more fine-grained sets.

Table 6.11 gives the results for these experiments. Increases and decreases for a given

Table 6.11: Results for neutral-polar feature ablation experiments

BoosTexter	Acc	Polar F	Neut F	**Ripper**	Acc	Polar F	Neut F
PARTS-OF-SPEECH	+	–	+	PARTS-OF-SPEECH	+++	+++	– – –
RELIABILITY-CLASS	+	–	+	RELIABILITY-CLASS	+++	+++	+
PRECEDED-POS	nc	–	nc	PRECEDED-POS	–	–	–
INTENSIFY	-	nc	-	INTENSIFY	–	– – –	–
RELCLASS-MOD	+	++	+	RELCLASS-MOD	+	+++	+
POLARITY-MOD	nc	–	+	POLARITY-MOD	–	+++	–
STRUCTURE	–	– – –	+	STRUCTURE	–	+	–
CURSENT-COUNTS	+	– – –	+	CURSENT-COUNTS	– –	+++	– – –
PNSENT-COUNTS	+	– – –	+	PNSENT-COUNTS	– – –	+++	– – –
CURSENT-OTHER	nc	–	+	CURSENT-OTHER	– – –	+++	– – –
TOPIC	+	+	+	TOPIC	–	+++	– – –
TiMBL	Acc	Polar F	Neut F	**SVM**	Acc	Polar F	Neut F
PARTS-OF-SPEECH	+	+++	+	PARTS-OF-SPEECH	– –	– – –	–
RELIABILITY-CLASS	+	+	nc	RELIABILITY-CLASS	+	–	+
PRECEDED-POS	nc	+	nc	PRECEDED-POS	nc	nc	nc
INTENSIFY	nc	nc	nc	INTENSIFY	nc	nc	nc
RELCLASS-MOD	+	+	+	RELCLASS-MOD	nc	+	nc
POLARITY-MOD	+	+	+	POLARITY-MOD	– –	– – –	– –
STRUCTURE	nc	+	–	STRUCTURE	–	+	–
CURSENT-COUNTS	–	+	–	CURSENT-COUNTS	–	–	–
PNSENT-COUNTS	+	+++	–	PNSENT-COUNTS	–	–	–
CURSENT-OTHER	+	+++	–	CURSENT-OTHER	–	–	–
TOPIC	–	+	–	TOPIC	–	–	–

Increases and decreases for a given metric as compared to the word+priorpol baseline are indicated by + or –, respectively; ++ or – – indicates the change is significant at the $p < 0.1$ level; +++ or – – – indicates significance at the $p < 0.05$ level; nc indicates no change.

metric as compared to the word+priorpol baseline (word baseline for Ripper) are indicated by + or –, respectively. Where changes are significant at the $p < 0.1$ level, ++ or – – are used, and where changes are significant at the $p < 0.05$ level, +++ or – – – are used. An "nc" indicates no change (a change of less than +/– 0.05) compared to the baseline.

What does Table 6.11 reveal about the performance of various feature sets for neutral-polar classification? Most noticeable is that no individual feature sets stand out as strong performers. The only significant improvements in accuracy come from the PARTS-OF-SPEECH and RELIABILITY-CLASS feature sets for Ripper. The significant improvements for Ripper are perhaps not surprising given that the Ripper baseline was much lower to begin with. Very few feature sets show any kind improvement for SVM. Again, this is not unexpected given that all the features together performed worse than the word+priorpol baseline for SVM. The performance of the feature sets for BoosTexter and TiMBL are perhaps the most revealing. In the previous experiments using all the features together, these algorithms produced classifiers with the same high performance. In these experiments, six different feature sets for each algorithm show improvements in accuracy over the baseline, yet none of those improvements are significant. This suggests that achieving the highest performance for neutral-polar classification requires a wide variety of features working together in combination.

I further tested this result by evaluating the effect of removing the features that produced either no change or a drop in accuracy from the respective all-feature classifiers. For example, I trained a TiMBL neutral-polar classifier using all the features except for those in the PRECEDED-POS, INTENSIFY, STRUCTURE, CURSENT-COUNTS, and TOPIC feature sets, and then compared the performance of this new classifier to the TiMBL, all-feature classifier. Although removing the non-performing features had little effect for boosting, performance did drop for both TiMBL and Ripper. The primary source of this performance drop was a decrease in polar recall: 2% for TiMBL and 3.2% for Ripper.

Although no feature sets stand out in Table 6.11 as far as giving an overall high performance, there are some features that consistently show improvements across the different algorithms. The reliability class of the clue instance (RELIABILITY-CLASS) improves accuracy over the baseline for all four algorithms. It is the only feature that does so. The

RELCLASS-MOD features give improvements for all metrics for BoosTexter, Ripper, and TiMBL, as well as improving polar F-measure for SVM. The PARTS-OF-SPEECH feature are also fairly consistent, improving proformance for all the algorithms except for SVM. There are also a couple of feature sets that consistently do not improve performance for any of the algorithms: the INTENSIFY and PRECEDED-POS features.

6.7.2 Polarity Classification

For the second step of recognizing contextual polarity, I classify the polarity of all clue instances identified as polar in step one. The features for polarity classification were listed above in Table 6.7 and described in Section 6.6.2

In this section, I investigate the performance of the polarity features under two conditions for step one: (1) perfect neutral-polar recognition and (2) automatic neutral-polar recognition. For condition 1, I identify the polar instances according to the gold-standard, manual contextual-polarity annotations. In the test data, 9,835 instances of the clues from the lexicon are polar in context according to the manual annotations. Experiments under condition 1 classify these instances as having positive, negative, or both (positive or negative) polarity. For condition 2, I take the best performing neutral-polar classifier for each algorithm, and use the output from those algorithms to identify the polar instances. Because polar instances are now being identified automatically, there will be noise in the form of misclassified neutral instances. Therefore, for experiments under condition 2 I include the neutral class and perform four-way classification instead of three-way. Condition 1 allows the investigation of the performance of the different polarity features without the noise of misclassified neutral instances. Also, because the set of polar instances being classified is the same for all the classification algorithms, condition 1 allows the performance of the polarity features to be compared across the algorithms. However, condition 2 is the more natural one. It shows how the noise of neutral instances affects the performance of the polarity features.

The sections below describe three sets of experiments. First, I investigate the performance of the polarity features used together for polarity classification under condition 1. As before, the word token and word+priorpol classifiers provide the baselines. In the second set

Table 6.12: Algorithm settings for polarity classification

Algorithm	Settings
BoosTexter	2000 rounds of boosting
TiMBL	k=1, MVDM distance metric
Ripper	-!s, -S 0.5
SVM	linear kernel

of experiments, I explore the performance of different sets of features for polarity classification, again assuming perfect recognition of the polar instances. Finally, I experiment with polarity classification using all the polarity features under condition 2, automatic recognition of the polar instances.

As I did for neutral-polar classification, I used the development data to select the settings for the algorithm parameters. The settings for the algorithms for polarity classification were selected based on the performance of the word+priorpol baseline classifier under condition 2. They are given in Table 6.12.

6.7.2.1 Classification Results: Condition 1 The results for polarity classification using all the polarity features, assuming perfect neutral-polar recognition for step one, are given in Table 6.13. For each algorithm, I give the results for the two baseline classifiers, followed by the results for the classifier trained using all the polarity features. For the metrics where the polarity features perform statistically better than both baselines (two-sided t-test, $p < 0.05$), the results are given in bold.

How well do the polarity features perform working all together? For all algorithms, the polarity classifier using all the features significantly outperforms both baselines in terms of accuracy, positive F-measure, and negative F-measure. These consistent improvements in performance across all four algorithms show that these features are quite useful for polarity classification.

One interesting thing that Table 6.13 reveals is that negative polarity words are much more straightforward to recognize than positive polarity words, at least in the MPQA corpus.

Table 6.13: Results for step 2 polarity classification using gold-standard polar instances

		Positive			Negative			Both		
	Acc	Rec	Prec	F	Rec	Prec	F	Rec	Prec	F
BoosTexter										
word token baseline	78.7	57.7	72.8	64.4	91.5	80.8	85.8	12.9	53.6	20.8
word+priorpol baseline	79.7	70.5	68.8	69.6	87.2	85.1	86.1	13.7	53.7	21.8
polarity features	**83.2**	**76.7**	74.3	**75.5**	89.7	**87.7**	**88.7**	11.8	54.2	19.4
TiMBL										
word token baseline	78.5	63.3	69.2	66.1	88.6	82.5	85.4	14.1	51.0	22.1
word+priorpol baseline	79.4	69.7	68.4	69.1	87.0	84.8	85.9	14.6	53.5	22.9
polarity features	**82.2**	**75.4**	**73.3**	**74.3**	88.5	**87.6**	**88.0**	18.3	34.6	23.9
Ripper										
word token baseline	70.0	14.5	74.5	24.3	98.3	69.7	81.6	09.1	74.4	16.2
word+priorpol baseline	78.9	75.5	65.2	70.0	83.8	86.4	85.1	09.8	75.4	17.4
polarity features	**83.2**	**77.8**	73.5	**75.6**	89.2	87.8	**88.5**	09.8	74.9	17.4
SVM										
word token baseline	69.9	62.4	69.6	65.8	76.0	84.1	79.9	14.1	31.2	19.4
word+priorpol baseline	78.2	76.7	63.7	69.6	82.2	86.7	84.4	09.8	75.4	17.4
polarity features	**81.6**	74.9	71.1	**72.9**	**88.1**	86.6	**87.3**	09.5	77.6	16.9

For the negative class, precisions and recalls for the word+priorpol baseline range from 82.2 to 87.2. For the positive class, precisions and recalls for the word+priorpol baseline range from 63.7 to 76.7. However, it is with the positive class that polarity features seem to help the most. With the addition of the polarity features, positive F-measure improves by 5 points on average; improvements in negative F-measures average only 2.75 points.

6.7.2.2 Feature Set Evaluation To evaluate the performance of the various features for polarity classification, I again performed a series of ablation experiments. As before, I started with the word+priorpol baseline classifier, added different sets of polarity features, trained new classifiers, and compared the results of the new classifiers to the word+priorpol baseline. Table 6.14 lists the sets of features tested in each experiment, and Table 6.15 shows the results of the experiments. Results are reported as they were previously in 6.7.1.2, with increases and decreases for a given metric as compared to the baseline indicated by + or –, respectively.

Table 6.14: Polarity feature sets for evaluation

Experiment	Features
NEGATION	negated, negated subject
POLARITY-MOD	modifies polarity, modified by polarity, conjunction polarity
SHIFTERS	general, negative, positive polarity shifters

Table 6.15: Results for polarity feature ablation experiments

:

		Positive			Negative		
	Acc	Rec	Prec	F	Rec	Prec	F
BoosTexter							
NEGATION	+++	++	+++	+++	+++	+	+++
POLARITY-MOD	++	+++	+	+++	+	++	+
SHIFTERS	+	+	+	+	+	+	+
TiMBL							
NEGATION	+++	+++	+++	+++	+++	+++	+++
POLARITY-MOD	+	+	+	+	–	+	+
SHIFTERS	+	+	+	+	–	+	+
Ripper							
NEGATION	+++	– –	+++	+++	+++	–	+++
POLARITY-MOD	+	+++	++	+++	+	+	+
SHIFTERS	+	–	+	+	+	–	+
SVM							
NEGATION	+++	–	+++	+++	+++	+	+++
POLARITY-MOD	+	–	+++	+	+	–	+
SHIFTERS	+	–	+	+	+	+	+

Increases and decreases for a given metric as compared to the word+priorpol baseline are indicated by + or –, respectively; ++ or – – indicates the change is significant at the $p < 0.1$ level; +++ or – – – indicates significance at the $p < 0.05$ level; nc indicates no change.

Table 6.16: Results for polarity classification without and with the *word token* feature

	Acc	Pos F	Neg F	Both F
BoosTexter				
excluding word token	82.5	74.9	88.0	17.4
all polarity features	83.2	75.5	88.7	19.4
TiMBL				
excluding word token	83.2	75.9	88.4	17.3
all polarity features	82.2	74.3	88.0	23.9
Ripper				
excluding word token	82.9	75.4	88.3	17.4
all polarity features	83.2	75.6	88.5	17.4
SVM				
excluding word token	81.5	72.9	87.3	16.8
all polarity features	81.6	72.9	87.3	16.9

Table 6.15 shows that all three sets of polarity features help to increase performance as measured by accuracy and positive and negative F-measures. This is true for all the classification algorithms. As might be expected, including the negation features has the most marked effect on the performance of polarity classification, with statistically significant improvements for most metrics across all the algorithms. The polarity modification features also seem to be important for polarity classification, in particular for disambiguating the positive instances. For all the algorithms except TiMBL, including the polarity modification features results in significant improvements for at least one of the positive metrics. The polarity shifters also help polarity classification, but including them does not result in significant improvements for any algorithm.

Another question that is interesting to consider is how much the *word token* feature contributes to the polarity classification results, given all the other polarity features. Is it enough to know the prior polarity of a word, whether it is being negated, and how it is related to other polarity influencers? To answer this question, I trained classifiers using all the polarity features except for the word token. Table 6.16 gives the results for these classifiers; for comparison, the results for the all-feature polarity classifiers are also given. Interestingly, excluding the *word token* feature produces only small changes in the overall

results. The results for BoosTexter and Ripper are slightly lower, while the results for SVM are practically unchanged. TiMBL actually shows a slight improvement, with the exception of the *both* category. This provides further evidence of the strength of the polarity features. Also, a classifier not tied to actual word tokens may potentially be a more domain independent classifier.

6.7.2.3 Classification Results: Condition 2

The experiments in Section 6.7.2.1 showed that the polarity features perform well under the ideal condition of perfect recognition of polar instances. The next question to consider is how well the polarity features perform under the more natural but less-than-perfect condition of automatic recognition of polar instances. To investigate this, the polarity classifiers (including the baselines) for each algorithm in these experiments start with the polar instances identified by the best performing neutral-polar classifier for that algorithm (from Section 6.7.1.1). The results for these experiments are given in Table 6.17. As before, statistically significant improvements over both baselines are given in bold.

How well do the polarity features perform in the presence of noise from misclassified neutral instances? The first observation comes from comparing Table 6.13 with Table 6.17: Polarity classification results are much lower for all classifiers with the noise of neutral instances. Yet in spite of this, the polarity features still produce classifiers that outperform the baselines. For three of the four algorithms, the classifier using all the polarity features has the highest accuracy. For BoosTexter and TiMBL, the improvements in accuracy over both baselines are significant. Also, for all algorithms, using the polarity features gives the highest positive and negative F-measures.

Because the set of polarity instances being classified by each algorithm is different, it is not possible to directly compare the results from one algorithm to the next.

6.7.3 Two-step versus One-step Recognition of Contextual Polarity

Although the two-step approach to recognizing contextual polarity allowed for a focused investigation of the performance of features for both neutral-polar classification and polarity

Table 6.17: Results for step 2 polarity classification using automatically identified polar instances

		Positive			Negative			Both			Neutral		
	Acc	R	P	F	R	P	F	R	P	F	R	P	F
BoosTexter													
word token	61.5	62.3	62.7	62.5	86.4	64.6	74.0	11.4	49.3	18.5	20.8	44.5	28.3
word+priorpol	63.3	70.0	57.9	63.4	81.3	71.5	76.1	12.5	47.3	19.8	30.9	47.5	37.4
polarity feats	**65.9**	73.6	62.2	**67.4**	84.9	72.3	**78.1**	13.4	40.7	20.2	31.0	50.6	**38.4**
TiMBL													
word token	60.1	68.3	58.9	63.2	81.8	65.0	72.5	11.2	39.6	17.4	21.6	43.1	28.8
word+priorpol	61.0	73.2	53.4	61.8	80.6	69.8	74.8	12.7	41.7	19.5	23.0	44.2	30.3
polarity feats	**64.4**	75.3	58.6	**65.9**	81.1	73.0	**76.9**	16.9	32.7	22.3	**32.1**	**50.0**	**39.1**
Ripper													
word token	54.4	22.2	69.4	33.6	95.1	50.7	66.1	00.0	00.0	00.0	21.7	76.5	33.8
word+priorpol	51.4	24.0	71.7	35.9	97.7	48.9	65.1	00.0	00.0	00.0	09.2	75.8	16.3
polarity feats	54.8	**38.0**	67.2	**48.5**	95.5	52.7	67.9	00.0	00.0	00.0	14.5	66.8	23.8
SVM													
word token	64.5	70.0	60.9	65.1	70.9	74.9	72.9	16.6	41.5	23.7	53.3	51.0	52.1
word+priorpol	62.8	89.0	51.2	65.0	88.4	69.2	77.6	11.1	48.5	18.0	02.4	58.3	04.5
polarity feats	64.1	90.8	53.0	66.9	**90.4**	70.1	79.0	12.7	52.3	20.4	02.2	61.4	04.3

Table 6.18: Results for contextual polarity classification for both two-step and one-step approaches

	Acc	Pos F	Neg F	Both F	Neutral F
BoosTexter					
2-step	74.5	47.1	57.5	12.9	83.4
1-step all feats	74.3	49.1	59.8	14.1	82.9
1-step –neut-pol feats	73.3	48.4	58.7	16.3	81.9
TiMBL					
2-step	74.1	47.6	56.4	13.8	83.2
1-step all feats	73.9	49.6	59.3	15.2	82.6
1-step –neut-pol feats	72.5	49.5	56.9	21.6	81.4
Ripper					
2-step	68.9	26.6	49.0	00.0	80.1
1-step all feats	69.5	30.2	52.8	14.0	79.4
1-step –neut-pol feats	67.0	28.9	33.0	11.4	78.6
SVM					
2-step	73.1	46.6	58.0	13.0	82.1
1-step	71.6	43.4	51.7	17.0	81.6

classification, the question remains: How does the two-step approach compare to recognizing contextual polarity in a single classification step? The results shown in Table 6.18 help to answer this question. The first row in Table 6.18 for each algorithm shows the combined result for the two stages of classification. For BoosTexter, TiMBL, and Ripper, this is the combination of results from using all the neutral-polar features for step one, together with the results from using all of the polarity features for step two[4]. For SVM, this is the combination of results from the word+priorpol baseline from step one, together with results for using all the polarity features for step two. Recall that the word+priorpol classifier was the best neutral-polar classifier for SVM (see Table 6.9). The second rows for BoosTexter, TiMBL, and Ripper show the results of a single classifier trained to recognize contextual polarity using all the neutral-polar and polarity features together. For SVM, the second row shows the results of classifying the contextual polarity using just the word token feature. This classifier outperformed all others for SVM.

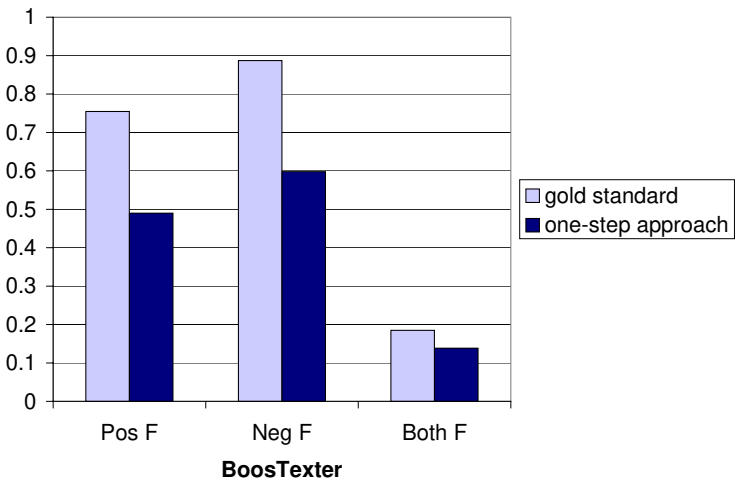
[4]To clarify, Section 6.7.2.3 only reported results for the polar instances identified in step one. Here, results are reported for all clue instances, including the correct and incorrect neutral classifications from step one.

When comparing the two-step and one-step approaches (the first two rows for each classifier), contrary to my expectations, the one-step approach performs about as well as the two-step approach for recognizing contextual polarity. For SVM, the difference in accuracy is significant, but this is not true for the other algorithms. One fairly consistent difference between the two approaches across all the algorithms is that the two-step approach always gives slightly higher neutral F-measure, and the one-step approach achieves higher F-measures for the polarity classes. The difference in negative F-measure is significant for BoosTexter, TiMBL, and Ripper. The exception to this is SVM. for SVM, the two-step approach achieves significantly higher positive and negative F-measures.

One last question to consider is how much the neutral-polar features contribute to the performance of the one-step classifiers. The third line in Table 6.18 for BoosTexter, TiMBL, and Ripper gives the results for a one-step classifier trained without the neutral-polar features. Although the differences are not always large, excluding the neutral-polar features consistently degrades performance in terms of accuracy and positive, negative, and neutral F-measures. The drop in negative F-measure is significant for all three algorithms, the drop in neutral F-measure is significant for BoosTexter and TiMBL, and the drop in accuracy is significant for TiMBL and Ripper (and for BoosTexter at the $p < 0.1$ level).

The modest drop in performance caused by excluding the neutral-polar features in the one-step approach may lead to the conclusion that discriminating between neutral and polar instances is helpful but not necessarily crucial. However, consider Figure 6.3. This figure shows the F-measures for the *positive*, *negative* and *both* classes for the BoosTexter classifier that uses the gold-standard neutral/polar instances (from Table 6.13) and for the BoosTexter one-step classifier that uses all features (from Table 6.18). Plotting the same sets of results for the other three algorithms produces very similar figures. The difference when the classifiers have to contend with the noise from neutral instances is dramatic. Although Table 6.18 shows that there is room for improvement across all the contextual polarity classes, Figure 6.3 shows that perhaps the best way to achieve these improvements is to improve the ability to discriminate the neutral class from the others.

Figure 6.3: Comparison of *positive*, *negative* and *both* class F-measures for the BoosTexter polarity classifier that uses the gold-standard neutral/polar classes and the BoosTexter one-step polarity classifier that uses all the features



6.8 RELATED WORK

There is a great deal of research in automatic sentiment analysis, ranging from work on learning the prior polarity (semantic orientation) of words and phrases to work characterizing the sentiment of documents. In this section, I review only the most closely related research, mainly, research in sentiment analysis that uses similar features or has similar findings, and other research in recognizing contextual polarity.

Identifying prior polarity (e.g., Hatzivassiloglou and McKeown (1997), Esuli and Sebastiani (2005), and Takamura et al. (2005)) is a different task than recognizing contextual polarity, although the two tasks are complementary. Whereas the goal of identifying prior polarity is to automatically acquire the polarity of words or phrases for listing in a lexicon, this research on recognizing contextual polarity begins with a lexicon of words with established prior polarities, and disambiguates the polarity being expressed by the phrases in the corpus in which instances of those words appear. To make the relationship between that task and this one clearer, some word lists that are used to evaluate methods for recognizing

prior polarity (positive and negative word lists from the General Inquirer (Stone et al., 1966) and lists of positive and negative adjectives created for evaluation by Hatzivassiloglou and McKeown (1997)) are included in the prior-polarity lexicon used in the experiments in this chapter.

For the most part, the features explored in this work differ from the ones used to identify prior polarity, with just a few exceptions. Using a feature to capture conjunctions between polarity clues was motivated in part by the work of Hatzivassiloglou and McKeown (1997). They use constraints on the co-occurrence in conjunctions of words with similar or opposite polarity for predicting the prior polarity of adjectives. Esuli and Sebastiani (2005) consider negation in some of their experiments involving WordNet glosses. Takamura et al. (2005) use negation words and phrases, including phrases such as *lack of* that are members in the lists of polarity shifters, and conjunctive expressions that they collect from corpora. Esuli and Sebastiani (2006) is also the only work in prior-polarity identification to include a *neutral* (*objective*) category and to consider a three-way classification between positive, negative, and neutral words. Although identifying prior polarity is a different task, they report a finding similar to mine, namely, that accuracy is lower when neutral words are included.

Some researchers in document-level sentiment classification have reported findings that are similar to some of the findings in this chapter. Bai et al. (2005) argue that dependencies among key sentiment terms are important for classifying document sentiment. Similarly, I found that features for capturing when clue instances modify each other are important for phrase-level classification, in particular, for identifying positive expressions. Gamon (2004) achieves his best results for document classification using a wide variety of features, including rich linguistic features, such as features that capture constituent structure, features that combine part-of-speech and semantic relations (e.g., sentence subject or negated context), and features that capture tense information. Similarly, the best results for phrase-level classification were achieved using a wide variety of features, many of which are linguistically rich. Kennedy and Inkpen (2006), report consistently higher results for document sentiment classification when select polarity influencers, including negators and intensifiers, are included[5].

[5] (Das and Chen, 2001; Pang, Lee, and Vaithyanathan, 2002; Dave, Lawrence, and Pennock, 2003) also represent negation. In their experiments, words which follow a negation term are tagged with a negation marker, and then treated as new words. Pang, Lee and Vaithyanathan report that representing negation in

Koppel and Schler (2006) demonstrate the importance of neutral examples for document-level classification. In this work, I show that being able to correctly identify neutral instances is also very important for phrase-level sentiment analysis.

Morinaga et al. (2002), Yu and Hatzivassiloglou (2003), Kim and Hovy (2004), Hu and Liu (2004), and Grefenstette et al. (2004)[6] have all worked on classifying the sentiment of sentences. They all begin, as I did, by first creating prior-polarity lexicons. Yu and Hatzivassiloglou assign a sentiment to a sentence by averaging the prior semantic orientations of instances of lexicon words in the sentence. Thus, they do not identify the contextual polarity of individual phrases containing clues, which is the focus of this work. Morinaga et al. only consider the positive or negative clue in each sentence that is closest to some target reference; Kim and Hovy, Hu and Liu, and Grefenstette et al. multiply or count the prior polarities of clue instances in the sentence. These researchers also consider local negation to reverse polarity, with Morinaga et al. also taking into account the negating effect of words like *insufficient*. However, they do not use the other types of features that I consider in my experiments. Kaji and Kitsuregawa (2006) take a different approach to recognizing positive and negative sentences. They bootstrap from information easily obtained in "Pro" and "Con" HTML tables and lists, and from one high-precision linguistic pattern, to automatically construct a large corpus of positive and negative sentences. They then use this corpus to train a Naive Bayes sentence classifier. In contrast to the research in this chapter, sentiment classification in all of the above research is restricted to identifying only *positive* and *negative* sentences (excluding the *both* and *neutral* categories). In addition, only one sentiment is assigned per sentence. The automatic systems that I developed assign contextual polarity to individual expressions, which allows for a sentence to be assigned to multiple sentiment categories. As my explorations of the contextual polarity annotations showed, it is not uncommon for sentences to contain more than one sentiment expression.

A few researchers in sentiment analysis have worked on classifying the contextual polarity of sentiment expressions (Yi et al., 2003; Popescu and Etzioni, 2005; Suzuki, Takamura, and

this way slightly helps their results, while Dave, Lawrence, and Pennock report a slightly detrimental effect. Whitelaw et al. (2005) also represent negation terms and intensifiers. However, in their experiments, the effect of negation is not separately evaluated, and intensifiers are not found to beneficial.

[6] In (Grefenstette et al., 2004), the units that are classified are fixed windows around named entities rather than sentences.

Okumura, 2006). This is the research most closely related to the research in this chapter. Yi et al. use a lexicon and manually developed patterns to classify contextual polarity. Their patterns are high-quality, yielding quite high precision over the set of expressions that they evaluate. Popescu and Etzioni use an unsupervised classification technique called **relaxation labelling** (Hummel and Zucker, 1983) to recognize the contextual polarity of words that are at the heads of select opinion phrases. They take an iterative approach, using relaxation labelling first to determine the contextual polarities of the words, then again to label the polarities of the words with respect to their targets. A third stage of relaxation labelling then is used to assign final polarities to the words, taking into consideration the presence of other polarity terms and negation. As I do, Popescu and Etzioni use features that represent conjunctions and dependency relations between polarity words. Suzuki et al. use a bootstrapping approach to classify the polarity of tuples of adjectives and their target nouns in Japanese blogs. Included in the features that they use are the words that modify the adjectives and the word that the adjective modifies. They consider the effect of a single negation term, the Japanese equivalent of *not*.

The research in this chapter differs from the above research on expression-level sentiment analysis in several ways. First, the set of expressions they evaluate is limited either to those that target specific items of interest, such as products and product features, or to tuples of adjectives and nouns. In contrast, I seek to classify the contextual polarity of all instances of the words in a large lexicon of subjectivity clues that appear in the corpus. Included in the lexicon are not only adjectives, but nouns, verbs, adverbs, and even modals. This work also differs from other research in the variety of features that I use. As other researchers do, I consider negation and the words that directly modify or are modified by the expression being classified. However, with negation, I have features for both local and longer-distance types of negation, and I take care to count negation terms only when they are actually being used to negate, excluding, for example, negation terms when they are used in phrases that intensify (e.g., *not only*). I also include contextual features to capture the presence of clues in the surrounding sentences, and features that represent the reliability of clues from the lexicon.

Finally, a unique aspect of the research in this chapter is the evaluation of different

features for recognizing contextual polarity. I evaluate not only features for discriminating between positive and negative polarity, but features for determining when a word is or is not expressing a sentiment in the first place (*neutral in context*). This is also the first work to evaluate the effect of neutral instances on the performance of features for discriminating between positive and negative contextual polarity.

6.9 CONCLUSIONS

In the research presented in this chapter, I tackled the problem of determining the contextual polarity of words and phrases, and showed that it is a much more complex problem than simply determining whether a word or phrase is positive or negative. In my analysis of the contextual polarity annotations in the MPQA Corpus, I found that positive and negative words from a lexicon are used in neutral contexts much more often than they are used in expressions of the opposite polarity. The importance of identifying when contextual polarity is neutral was further revealed in my classification experiments: When neutral instances are excluded, the performance of features for distinguishing between positive and negative polarity greatly improves.

A focus of this chapter was on understanding which features are important for recognizing contextual polarity. I experimented with a wide variety of linguistically-motivated features, and I evaluate the performance of these features using several different machine learning algorithms. Features for distinguishing between neutral and polar instances were evaluated, as well as features for distinguishing between positive and negative contextual polarity. For classifying neutral and polar instances, I found that, although some features produced significant improvements over the baseline in terms of polar or neutral recall or precision, it was the combination of all features together that was needed to achieve significant improvements in accuracy. For classifying positive and negative contextual polarity, features for capturing negation proved to be the most important. However, I found that features also performed well that capture when a word is (or is not) modifying or being modified by other polarity terms.

7.0 REPRESENTING ATTITUDES AND TARGETS

Private states in language are often quite complex in terms of the attitudes they express and the targets of those attitudes. For example, consider the private state represented by the direct subjective phrase "are happy" in the following sentence.

(7.1) "I think people are happy because Chavez has fallen."

In this sentence, the word "happy" expresses a positive attitude, specifically, the positive sentiment of the people toward the fall of Chavez. However, the private state attributed to the people in this sentence encompasses more than just a positive sentiment. There is a second attitude, a negative sentiment toward Chavez himself, which can be inferred from the phrase "happy because Chavez has fallen."

Just as a private state may involve more than one type of attitude, an attitude may be directed toward more than one target. In sentence (7.2) there is a private state being expressed by Tsvangirai.

(7.2) Tsvangirai said the election result was a clear case of highway robbery by Mugabe, his government and his party, Zanu-PF.

The negative sentiment of this private state is expressed with the phrase "a clear case of highway robbery," and it is directed toward two things: "the election results" and "Mugabe, his government and his party, Zanu-PF."

In this chapter, I extend the original conceptual representation of private states (Chapter 3) to better model attitudes and their targets. In the original conceptualization, attitudes are represented with the *attitude type* attribute in direct subjective and expressive subjective element frames, and targets are represented with the *target* attribute in direct subjective

frames. A drawback to representing attitudes and targets in this way is that it does not allow for multiple attitudes and targets to be associated with a private state. In the new representation, attitudes and targets are conceptualized as annotation frames, with target frames linking to attitude frames and attitude frames linking to private state frames. This representation gives the flexibility needed to associate multiple attitudes and targets with a single private state.

The new representation also includes a new, more clearly-defined set of attitude types. What types of attitudes are useful for NLP is an open question and, at least to a certain extent, application dependent. Sentiment, which so far has received the most attention, is clearly important. However, as the contextual polarity annotations in Chapter 6 showed, sentiment is far from the only type of attitude: 45% of the subjective expressions in the MPQA corpus express some type of attitude other than a sentiment. Sentence 7.1 above contains an example of a private state expressing a type of attitude other than sentiment. In the context of Sentence 7.1, the word "think" is being used to express an opinion about what is true according to its source. I developed the set of attitude types presented in this chapter with an eye toward what would be useful for NLP applications, in particular an application like question answering. I hypothesize that these attitude types can be reliably annotated, and that they will provide provide good coverage of the private states expressed in the MPQA Corpus.

With my extension to the conceptualization, I aim to improve on one more aspect of the representation of private states: intensity. Judging the intensity of private states is challenging. In Chapter 3, I evaluated inter-annotator agreement for the various intensity judgments in the original conceptualization. Although inter-annotator agreement was acceptable for the *intensity* and *expression-intensity* of the combined set of direct subjective and objective speech event annotations, agreement for the intensity of expressive subjective elements was low. One way in which intensity judgments might be improved is to judge intensity with respect to attitude type, for example, to compare the intensity of a positive sentiment to other positive sentiments rather than very dissimilar types of attitudes such as speculations or intentions. Thus, in the new conceptualization, I define intensity explicitly according to the new set of attitude types.

7.1 CONCEPTUAL REPRESENTATION

In this section, I describe my extensions to the conceptual representation for attitude types and targets. I begin by introducing the new set of attitude types, and then describe the new attitude and target annotation frames and how they are integrated into the overall conceptual representation for private states. At the end of the section I give a graphical example to illustrate the new annotations.

7.1.1 Types of Attitude

When determining a set of attitude types, there are any number of possible distinctions that might be considered. The attitude types in the original conceptualization distinguish very generally between *positive* and *negative* attitudes, with other types of attitudes being lumped together into one category. For the new set of attitude types, my goal is to define more fine-grained distinctions. However, some distinctions may actually be too fine grained to be of use to an application. Would an application such as question answering benefit from being able to distinguish between a positive emotion and a positive evaluation? Or, would distinguishing between positive sentiments (which include both emotions and evaluations) and intentions be more helpful? Working with the annotators of the MPQA Corpus, looking at the private states already annotated, and keeping in mind what might be useful for an application like QA, I developed the set of attitude types listed in Table 7.1.

At the coarser-level of distinction, there are six attitude types: sentiment, agreement, arguing, intention, speculation, and all other attitudes. Sentiment, agreement, arguing, and intention may be further broken down into positive and negative variants. Below I define and give examples of each of the attitude types and their targets. In each example, the span of text where the attitude is expressed is in bold, and the span of text that denotes the target of the attitude (if a target is given) is in angle brackets.

7.1.1.1 Sentiments *Sentiments* are positive and negative emotions, evaluations, and stances. This is the same definition of sentiment that was used in Chapter 6. The target of

Table 7.1: Set of attitude types

Sentiment	**Agreement**
Positive Sentiment	Positive Agreement
Negative Sentiment	Negative Agreement
Arguing	**Intention**
Positive Arguing	Positive Intention
Negative Arguing	Negative Intention
Speculation	**Other Attitude**

a sentiment is what the sentiment is directed toward. Sentence 7.3 contains an example of a positive sentiment, and Sentence 7.4 contains an example of a negative sentiment.

Positive Sentiment:
(7.3) The Namibians went as far as to say ⟨Zimbabwe's election system⟩ was "**water tight, without room for rigging**".

Negative Sentiment:
(7.4) His disenfranchised supporters **were seething**.

7.1.1.2 Agreement Private states in which a person does or does not agree, concede, consent, or in general give assent to something fall into the category of *Agreement*. Agreement includes both agreeing with a statement or idea and agreeing to an action. The target for this attitude type is what is (or is not) being agreed to. Sentence 7.5 gives an example of positive agreement, and sentence 7.6 gives an example of a negative agreement. Sentence 7.7 has examples of both negative ("differed over") and positive ("agreed") agreement.

Positive Agreement:
(7.5) Republicans **concede** that ⟨at this point it could be his only option⟩.

Negative Agreement:
(7.6) Afghanistan is now under US bombardment for **refusing** ⟨to hand over the chief suspect in the Sept. 11 attacks on New York and Washington⟩.

(7.7) Japanese Prime Minister Junichiro Koizumi and visiting U.S. president George W. Bush **differed over** ⟨the Kyoto Protocol and how to prevent global warming⟩ but **agreed** ⟨to cooperate on that issue⟩.

7.1.1.3 Arguing Private states in which a person is arguing or expressing a belief about what is true or should be true in his or her view of the world are categorized as *Arguing*. Arguing attitudes include private states where the source is arguing for or against something.

Deciding on what spans to annotate for arguing attitudes (and speculation (Section 7.1.1.5), which is similar to arguing) and their targets actually turned out to be a challenging part of the annotation scheme development. In initial annotation rounds with another annotator, there was a great deal of inconsistency in what spans were marked for arguing attitudes and their targets, even though there was agreement that arguing was present. Eventually, I decided on the following strategy for marking arguing attitude and arguing target spans, because it seemed to produce the most consistent span annotations: mark the arguing attitude on the span of text expressing the argument or *what the argument is*, and mark *what the argument is about* as the target of the arguing attitude.

Sentences 7.8 and 7.9 contain examples of positive arguing attitudes, and sentence 7.10 and 7.11 contain negative arguing attitudes.

Positive Arguing:
(7.8) Iran **insists ⟨its nuclear program⟩ is purely for peaceful purposes**.

(7.9) Putin remarked that ⟨the events in Chechnia⟩ "**could be interpreted only in the context of the struggle against international terrorism**."

Negative Arguing:
(7.10) Officials in Panama **denied that ⟨Mr. Chavez or any of his family members⟩ had asked for asylum**.

(7.11) "⟨It⟩ **is analogous to the US crackdown on terrorists in Afghanistan**," Ma said.

7.1.1.4 Intentions *Intentions* include aims, goals, plans, and other *overt* expressions of intention. Positive intentions are straightforward. Negative intentions are the opposite of positive intentions. They are the intentions that the source of the private state is described explicitly as **not** holding. The target of an intention is the thing that is (or is not) the aim,

goal, plan, or intention. Sentence 7.12 has an example of a positive intention, and the private state in sentence 7.13 is an example of a negative intention.

Positive Intention:
(7.12) The Republic of China government believes in the US **commitment** ⟨to separating its anti-terrorism campaign from the Taiwan Strait issue⟩, an official said Thursday.

Negative Intention:
(7.13) The Bush administration **has no plans** ⟨to ease sanctions against mainland China⟩.

7.1.1.5 Speculations Private states in which a person is speculating about what is or is not true, or what may or may not happen, are categorized as *Speculation*. Similar to arguing, the span of text marked for speculation is *what the speculation is*, and *what the speculation is about* is marked as the target of the speculation. Sentence 7.14 gives an example.

(7.14) ⟨The president⟩ is **likely to endorse the bill**.

7.1.1.6 Other Attitudes The hope is that most private states will fall into the set of attitudes described above. However, for those that do not there is this category. Private states that would be captured by this catch-all category are neutral emotions (emotions that don't seem clearly positive or negative), cognition, and general uncertainty. Sentences 7.15 and 7.16 give two examples of other attitudes.

(7.15) To **the surprise of** many, ⟨the dollar hit only 2.4 pesos and closed at 2.1⟩.

(7.16) "I'm **not sure** whether ⟨I should wait in line or sell to one of the street traders⟩," said Fabian, a 36-year old attorney.

In sentence 7.15, it is not clear from the context whether the emotion surprise is positive or negative, so it is categorized as other. In sentence 7.16, Fabian, the source of the private state, is expressing his uncertainty.

7.1.2 Attitude Frames

When considering the representation of the new attitude annotations, the major question to address, aside from what set of attitude types to use, is to which spans of text should

the new attitudes be anchored? Arguing and speculation attitude types created their own challenge, and the spans to mark for these attitudes were described above. Sentence 7.17 illustrates a more general problem.

> (7.17) The MDC leader said systematic cheating, spoiling tactics, rigid new laws, and shear obstruction – as well as political violence and intimidation – were just some of the irregularities practised by the authorities in the run-up to, and during the poll.

In this sentence, there are five private state frames attributed to the MDC leader: a direct subjective frame anchored to "said," and four expressive subjective element frames anchored respectively to "systematic cheating . . . obstruction," "as well as," "violence and intimidation," and " just some of the irregularities." One option is to create an attitude frame for each of the private state frames. However, this would be very redundant, both in the expressions that would be annotated and in the sentiment annotations that would result. A better solution is to annotate the span of text that expresses the attitude of the overall private state represented by the direct subjective frame. Specifically, for each direct subjective frame, first the attitude type(s) being expressed by the source of the direct subjective frame are determined by considering the text anchor of the frame and everything within the scope of the annotation attributed to the source. Then, for each attitude type identified, an attitude frame is created and anchored to whatever span of text completely captures the attitude type. In sentence 7.17, this results in just one attitude frame being created to represent the negative attitude of the MDC leader. The anchor for this attitude frame begins with "systematic cheating" and ends with "irregularities."

To tie the attitude frames back to the direct subjective frame, each attitude annotation is given a unique, alphanumeric identifier, and a new *attitude link* attribute is created in the direct subjective frame. The value of the *attitude link* attribute is a list of one or more attitude frame identifiers.

Figure 7.1 gives the attributes for the attitude frame. The *id* attribute is the unique identifier used to link the attitude frame back to its corresponding direct subjective frame. The *text anchor* attribute points to the span of text on which the attitude is marked. The *attitude type* attribute is one of the attitude types described in the previous section. The *target link* attribute is used to link the attitude frame to its targets. For the rare cases in

Figure 7.1: Attitude frame

- **id:** a unique, alphanumeric ID for identifying the attitude annotation. The ID is used to link the attitude annotation to the private state that it is a part of.
- **text anchor:** a pointer to the span of text that captures the attitude being expressed.
- **attitude type:** type of attitude being expressed (Table 7.1).
- **target link:** list of one or more target frame IDs, or the string *none*.
- **intensity:** *low*, *low-medium*, *medium*, *medium-high*, *high*, *high-extreme*.
- **properties:**
  - **inferred:** true, if the attitude is inferred. (The *inferred* property will be described in more detail in Section 7.1.5.)
  - **sarcastic:** true, if the attitude is realized through sarcasm. (This attribute is discussed in more detail in Section 7.1.5.)
  - **repetition:** true, if the attitude is realized through the repetition of words, phrases, or syntax. (This attribute is discussed in more detail in Section 7.1.5.)
  - **contrast:** true, if the attitude is realized only through contrast with another attitude. (This attribute is discussed in more detail in Section 7.1.5.)

which an attitude has no clear target, the *target link* attribute is assigned the string *none*. Otherwise, it is a list of one or more target frame IDs (described in Section 7.1.3). The *intensity* attribute captures the intensity of the attitude type being expressed. Table 7.2 defines how the intensity for each attitude type should be evaluated. In addition to defining intensity more explicitly and with respect to the different attitude types, the values for intensity are more fine-grained than those used for the various intensity attributes in the original conceptualization. The hope is that this as well will help to improve inter-annotator agreement for intensity judgments. The remaining properties of the attitude frame are described later in Section 7.1.5.

7.1.3 Target Frames

Once an attitude frame has been created, the span of text representing the target of the attitude is identified and a target frame is created and anchored to that text span. Like the attitude frames, each target frame is given a unique, alphanumeric identifier. These identifiers and the *target link* attributes on the attitude frames are used to tie the attitude

Table 7.2: Measures of intensity for different attitude types

Attitude Type	**Measure of Intensity**	**Example**
Positive Sentiment	degree of positiveness	*like* < *love*
Negative Sentiment	degree of negativeness	*criticize* < *excoriate*
Positive Agreement	degree of agreement	*mostly agree* < *agree*
Negative Agreement	degree of disagreement	*mostly disagree* < *completely disagree*
Positive Arguing	degree of certainty/strength of belief	*critical* < *absolutely critical*
Negative Arguing	degree of certainty/strength of belief	*should not* < *really should not*
Positive Intention	degree of determination	*promise* < *promise with all my heart*
Negative intention	degree of determination	*no intention* < *absolutely no intention*
Speculation	degree of likelihood	*might win* < *really might win*

and target frames together. The target frame is given in Figure 7.2.

7.1.4 Example

To help to illustrate the new attitude and target frames and how they fit in with the original conceptual representation, Figure 7.3 gives the various direct subjective, attitude, and target frames for sentence 7.18 and shows how they are all linked together.

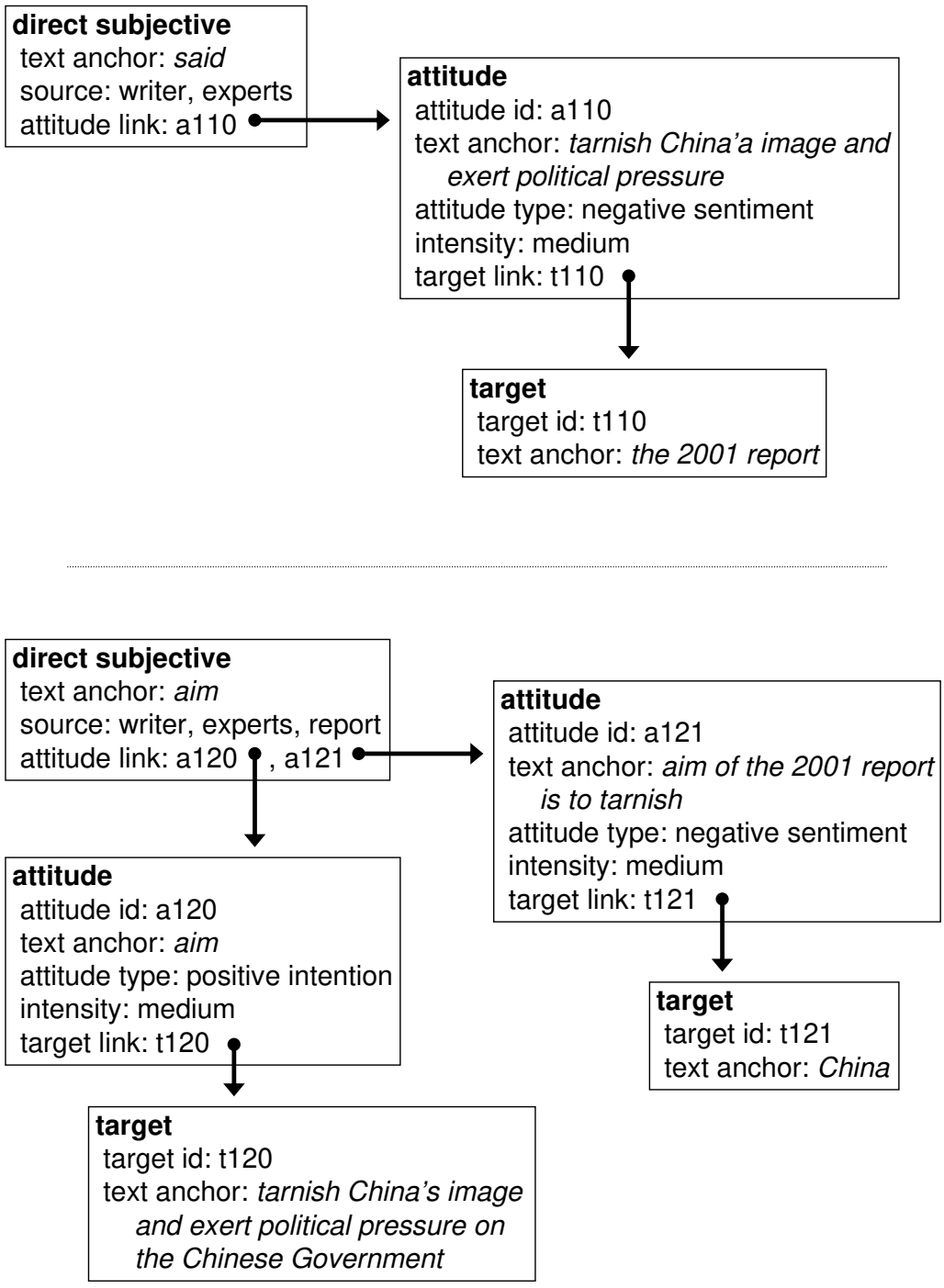
> (7.18) Its aim of the 2001 report is to tarnish China's image and exert political pressure on the Chinese Government, human rights experts said at the seminar held by the China Society for Study of Human Rights (CSSHR) on Friday.

There are two direct subjective frames in sentence 7.18. The private state represented by the direct subjective frame for "said" has one attitude with one target. The attitude is a negative sentiment expressed by the phrase "tarnish China's image and exert political

Figure 7.2: Target frame

- **id:** a unique alphanumeric ID for identifying the target annotation. The ID is used to link the target to the attitude frame.
- **text anchor:** a pointer to the span of text that denotes the target.

Figure 7.3: Private state, attitude, and target frames for sentence 7.18



pressure." The target of the negative sentiment is the 2001 report.

There are two attitudes for the private state represented by the direct subjective frame for "aim." The first attitude is a positive intention with the target tarnishing China's image and exerting political pressure on the Chinese Government. The second attitude is the negative sentiment that is conveyed by the phrase "aim of the 2001 report is to tarnish." Having an aim to tarnish indicates a negative sentiment. The target of the negative sentiment is China.

7.1.5 Additional Characteristics of Attitudes

The attitude frame has four additional properties that are used to mark particular characteristics of attitudes when they are relevant. The first of these properties is for marking when an attitude is *inferred*. The remaining properties represent characteristics of how attitudes are sometimes expressed. I include these properties in the attitude frame because I feel they may be useful in developing automatic systems for recognizing different types of attitudes.

7.1.5.1 Inferred Attitudes Most attitudes are directly evoked by the words and phrases that are used to express a private state. However, sometimes attitudes are *inferred*. For example, in the sentence *I think people are happy because Chavez has fallen* (sentence 7.1 above), the negative sentiment of the people toward Chavez is an inferred attitude. The most prominent attitude of the private state attributed to the people is a positive sentiment toward Chavez's fall, but the negative sentiment toward Chavez is only a short inference away.

One problem with marking inferred attitudes is that it is very easy to start "digging too deep" and inferring any number of very subtle attitudes. To cut down on the possibilities for this, annotators were instructed to mark only inferred attitudes that have people or other entities as their targets.

7.1.5.2 Characteristics of How Attitudes Are Expressed The properties that represent various ways attitudes may be expressed are *sarcastic*, *repetition*, and *contrast*.

The *sarcastic* property is for marking attitudes expressed using sarcasm. In general, I

believe this property will be of interest for NLP applications working with opinions. Detecting sarcasm may also help a system learn to distinguish between positive and negative sentiments. The sarcasm in sentence 7.19 below makes the word "Great" an expression of negative rather than positive sentiment.

(7.19) "Great, keep on buying dollars so there'll be more and more poor people in the country," shouted one.

The *repetition* property is used when an attitude and its intensity are expressed at least in part using the repetition of a word or phrase within a sentence or within several consecutive sentences. In sentence 7.20, the repetition of the phrase "a window" contributes a great deal to the intensity of the positive sentiment expressed by Taiwan.

(7.20) Taiwan's WTO access has given Taiwan a window to the world, a window to the century and a window of opportunity ...

The *contrasted* property is used to mark positive and negative attitudes where the type of attitude is only evident because the attitude is contrasted with an attitude of the opposite polarity. For example, consider the attitudes for the Italian senator and the United States in sentence 7.21.

(7.21) The Italian senator's words are in sharp contrast to what was contained in the so-called China human rights report compiled by the United States, which blindly accuses China of restricting religious freedom in Tibet.

The negative sentiment of the United States toward China is clearly indicated by the phrase "blindly accuses China of restricting religious freedom in Tibet." However, the positive sentiment of the Italian senator is understood only because it is contrasted with the negative sentiment of the US. The attitude for the Italian senator would thus be marked with the *contrast* property.

7.2 AGREEMENT STUDIES

In this section, I test the general hypothesis that the extensions to the conceptual representation of private states presented in the previous section can be reliably annotated. Given documents already annotated with private state frames according to the original conceptual representation, I first evaluate whether annotators agree about the attitudes for these private states. I then turn to the question of whether judging intensity according to attitude type gives an improvement in intensity agreement. Finally, I evaluate how well annotators agree in their target frame annotations.

I conducted two inter-annotator agreement studies. In the first study, another annotator and I independently annotated 13 documents with 325 sentences and 409 direct subjective annotations. Two months later, during which we at times discussed our annotations, we annotated another 11 documents with 211 sentences and 207 direct subjective annotations. All intensity and contextual polarity attributes were removed from the existing private state annotations in these documents before each study began.

7.2.1 Agreement for Attitude Frames and Attitude Types

Measuring agreement for attitudes requires first identifying and aligning the sets of attitudes marked by both annotators. If each annotator marked only one attitude for a given direct subjective frame d, the process of matching up their attitude annotations is straightforward. Let d_a and d_b be the attitude frames marked by the two annotators for d. If the text anchors of d_a and d_b overlap, then d_a is said to **match** d_b, and the two attitude frames are included in the set of attitudes marked by both annotators.

When one annotator or both mark more than one attitude for d, the process of matching up the attitude frames is a bit more complicated, involving both the text anchors and the attitude types of the attitude frames. No attitude marked by annotator a is allowed to match with more than one attitude marked by annotator b, and vice versa. Thus, if annotator a marked two attitudes on d, d_{a1} and d_{a2}, and annotator b only marked one attitude, d_{b1}, and if both d_{a1} and d_{a2} overlap with d_{b1}, only one of a's annotations can be matched with d_{b1}. When

Table 7.3: Inter-annotator agreement: Attitudes

	$\|A\|$	$\|B\|$	$recall(a\|\|b)$	$recall(b\|\|a)$	F-measure
Study 1	515	549	0.91	0.86	0.88
Study 2	247	283	0.95	0.83	0.89

choosing which attitude of a to match to d_{b1}, preference is given first to whichever attitude has the same attitude type as d_{b1}, and then to whichever attitude of a has a text anchor with the larger overlap with the text anchor of d_{b1}. There are direct subjective annotations in which both annotators mark multiple attitudes that overlap. These are more complicated, but the matches are resolved in a similar way.

To evaluate how well the two annotators agreed on identifying the same set of attitude annotations, I use recall and F-measure, just as I did for measuring inter-annotator agreement for private state and speech event text anchors in Chapter 3. As before, recall is calculated with respect to each annotator. The recall of a with respect to b is

$$recall(a||b) = \frac{|A \text{ matching } B|}{|A|}$$

and the recall of b with respect to a is

$$recall(b||a) = \frac{|B \text{ matching } A|}{|B|}$$

Table 7.3 gives the agreement for the attitude frames marked by the annotators in the two studies. The first two columns in the table show the number of attitudes marked by each annotator in the two studies, followed by their respective recalls. The last column is F-measure. In study 1, the two annotators agree on a total of 470 attitude frames, and in study 2 they agree on 235 attitude frames. This results in an average F-measure of 0.885 over the two studies, which indicates that the annotators largely agree on the attitude frames that they marked.

Now that the set of attitudes marked by both annotators has been identified and the attitude annotations have been match up, I can use Cohen's Kappa (κ) to evaluate how well

Table 7.4: Inter-annotator agreement: Attitude-types

	Fine		Conflated	
	κ	%	κ	%
Study 1	0.79	83%	0.78	86%
Study 2	0.81	85%	0.77	86%

the annotators agree on the attitude-types of the attitudes that they marked. As Table 7.4 shows, agreement for attitudes-types is high, with κ-values near 0.80. The first two columns in the table give κ and percent agreement values for the finer-grained set of attitude types. The last two columns give κ and percent agreement for the more general set of attitude types, in which the positive and negative variants are conflated (e.g., positive sentiment and negative sentiment are conflated into one category).

One interesting finding from these studies is that very few of the disagreements come from the annotators agreeing about the general type, but disagreeing about the polarity. There are only two positive sentiment/negative sentiment disagreements, and one positive arguing/negative arguing disagreement in study 2. In study 1, only 12 (15%) of the disagreements are between positive and negative attitudes of the same general category. By far the majority of disagreements in attitude-type judgments are between different general categories. Table 7.5 gives the contingency table showing these disagreements for study 1.

7.2.2 Agreement for Attitude Intensity

To evaluate the inter-annotator agreement for the intensity of attitudes, I again use the sets of matched attitude frames identified by both annotators. As I did in Chapter 3, I use Krippendorff's α to calculate agreement for intensity. Like κ, Krippendorff's α takes into account chance agreement, but unlike κ, it can be used to calculate agreement for ordinal judgments.

With α, a distance metric is used to weight disagreements. When measuring agreement for the different intensity judgments that are part of the original conceptualization, I used

Table 7.5: Confusion matrix for conflated attitude-type agreement for Study 1

		B						
		Sentiment	Agreement	Arguing	Intention	Speculation	Other	Total
	Sentiment	**203**	3	13	2	0	13	234
	Agreement	0	**5**	2	0	0	2	9
	Arguing	6	1	**145**	0	0	4	156
A	Intention	1	1	0	**17**	0	3	22
	Speculation	1	0	0	0	**6**	2	9
	Other	4	2	3	3	0	**28**	40
	Total	215	12	163	22	6	52	470

the scale [0,1,2,3], where 0 represented neutral and 3 represented high. The scale that I use for the intensity of attitudes is more fine-grained, but it can still be matched to the original scale by mapping the *low-medium* and *medium-high* ratings to mid-points (1.5 and 2.5), and by merging the *high* and *high-extreme* ratings. With this numeric scale of intensity, I can use the square of the difference between any two disagreements as the distance metric. Thus, the distance weight is 0.25 for any disagreement that differs by one-half (e.g., low-medium and medium), the distance weight is 1 for any disagreement that differs by 1 (e.g., low and medium), the weight is 4 for any disagreement that differs by two (e.g., low and high).

The α-agreement scores for attitude intensity for the two agreement studies are 0.65 and 0.61. These values are not high. Krippendorff recommends a value of at least 0.67 in order to draw tentative conclusions about reliability.

I hypothesized that defining intensity according to attitude type and using a finer-grained intensity scale would result in better intensity agreement as compared to how intensity was judged in the original conceptualization. In the agreement study reported in Chapter 3, average pairwise α-agreement for the intensity of expressive subjective elements was 0.46, with the highest pairwise agreement being 0.52. I did not previously report agreement for the intensity of direct subjective frames [1]. However, by identifying the set of direct subjective

[1] Agreement for the intensity and expression-intensity of the combined set of direct subjective and objective speech event frames was reported, but these are not directly comparable.

frames marked by each annotator pair in the Chapter 3 study, I calculate that the average pairwise α for direct subjective intensity was 0.44, with the highest pairwise agreement being 0.56.

At first glance, α-agreement for attitude intensity is higher than these earlier intensity agreement scores. However, is the agreement higher because of how it was defined (according to attitude type), or because of the finer-grained intensity ratings? I experimented with different strategies for conflating the mid-point intensity ratings (e.g., low-medium) so that the rating scale for the attitude intensity would be exactly the same as the intensity scale used in the previous study. With low-medium merged with medium and medium-high merged with high, α-agreement drops to 0.59 for study 1 and 0.55 for study 2. These agreement scores are still higher than the α-agreement for the best annotator pair agreements in the earlier study, but not by much. Judging intensity according to attitude type may be helpful, but the current annotation study does not provide very strong evidence in support of that hypothesis.

7.2.3 Agreement for Targets

In this section, I evaluate the inter-annotator agreement for the targets of the attitude frames identified by both annotators. Recall that the *target-link* attribute marked on every attitude frame either has the value *none*, or it is a list of one or more target frame ids. For the purpose of measuring target agreement, I treat the *none* value as a special type of target. If two matching attitude frames both have targets that are *none*, then the targets of those attitudes are also a match. To calculate whether two targets (other than *none*) of two matching attitude frame do themselves match, I look only at whether the text anchors for the targets overlap. If the text anchors overlap, then the two targets match; otherwise, they do not match.

Unlike with attitudes, I do allow a target marked by one annotator to match with more than one target marked by the other annotator. This is fairly uncommon, but it does happen, for example, when one annotator chooses to break one large target span into two different targets. This is what happened with the targets marked for the attitude anchored

Table 7.6: Inter-annotator agreement: Targets

	Attitudes	$\|A\|$	$\|B\|$	$recall(a\|\|b)$	$recall(b\|\|a)$	F-measure
Study 1	470	479	503	0.85	0.85	0.85
Study 2	235	244	247	0.86	0.86	0.86

on "criticized" in Sentence 7.22. The targets marked for each annotator are in angle-brackets.

(7.22)
A: US Senate Majority leader Tom Daschle **criticized** on Monday ⟨President George W. Bush for his remarks that described Iran, Iraq and the Democratic People's Republic of Korea (DPRK) as "axis of evil"⟩.

B: US Senate Majority leader Tom Daschle **criticized** on Monday ⟨President George W. Bush⟩ for ⟨his remarks that described Iran, Iraq and the Democratic People's Republic of Korea (DPRK) as "axis of evil"⟩.

Because the annotators in essence are still capturing the same entities for the target of the attitude, I decided it was appropriate to allow multiple target matches.

Table 7.6 gives the results for target agreement. As for the attitude frames, target agreement is measured using F-measure and recall. The first column of the table lists the number of matching attitudes for each study. It is the targets marked on these attitudes that I consider when calculating target agreement. The next two columns give the number of targets marked on these attitudes by annotators a and b, followed by the recall for each annotator. Target agreement is very similar for both studies, around 0.85. Although lower than agreement for the attitude frames, the study shows that annotators largely agree on what are the targets of attitudes.

7.3 OBSERVATIONS

About two-thirds of the documents in the MPQA Corpus version 2.0 have been annotated with a layer of attitude and target annotations. In this section, I briefly explore what has

Table 7.7: Distribution of attitude types for attitude frames and direct subjective frames

Attitude Type	% of Attitude Frames	% of Direct Subjective Frames
Positive Sentiment	16.7	20.0
Negative Sentiment	32.8	38.0
Positive Agreement	1.9	2.4
Negative Agreement	1.7	2.1
Positive Arguing	25.6	30.1
Negative Arguing	6.4	7.7
Positive Intention	5.2	6.5
Negative Intention	0.5	0.6
Speculation	2.3	3.0
Other Attitude	7.8	9.8

actually been annotated in terms of various distributions of the attitude annotations in a set of 284 documents. I call this set of documents the **attitude dataset**. I use the attitude dataset in the next chapter in my experiments in automatic attitude recognition.

There are 4,499 sentences in the attitude dataset. Of these sentences, 2,829 (63%) are subjective (i.e., they contain at least one direct subjective frame), with a total of 4,538 direct subjective frames and 5,739 attitude annotations. This means that on average, there are 1.6 direct subjective frames and 2 attitude frames in every subjective sentence. The majority of direct subjective frames, 80%, are linked to just one attitude frame. 18% of direct subjective frames are linked to two attitudes, and a very small 2% are linked to three attitudes. There is one direct subjective frame linked to four attitudes.

Table 7.7 shows two distributions of attitude types. The first column gives the distribution of attitude types for all the attitude frames marked in the dataset. The second column gives the distribution of the direct subjective frames with respect to the types of attitudes they are linked to. Because a direct subjective frame can be linked to more than one type of attitude, these percentages will not sum to 100.

As Table 7.7 shows, sentiments and arguing attitudes make up the largest number of attitude types. Almost 50% of the attitude annotations are sentiments, and nearly one-third are arguing. Interestingly, of the remaining attitude types, other attitude is the category

with the next highest number of attitude. However, fewer than 10% of the attitudes are marked as other, showing that the set of attitudes proposed in this chapter do have fairly good coverage of the types of private states expressed in the news.

7.4 RELATED WORK

Research into types of attitudes and models of emotion has been the focus of work in linguistics and psychology for many years. In psychology, for example, there is a long tradition of using hand-compiled emotion lexicons in experiments to help develop or support various models of emotion. One line of research (e.g., Osgood et al. (1957), Heise (1965), Russell (1980), and Watson and Tellegen (1985)) uses factor analysis to determine dimensions for characterizing emotions. Dimensions corresponding to polarity and intensity are two that are consistently identified. Other researchers (e.g., de Rivera (1977), Ortony et al. (1987), and Johnson-Laird and Oatley (1989)) developed taxonomies of emotions. My goals in developing the set of attitude types presented in this chapter and the goals of these works in psychology and linguistics are quite different. I am not interested in building models or taxonomies of emotion, but rather in identifying types of attitude that would be useful to recognize for improving NLP systems.

Appraisal Theory (Martin, 2000; White, 2002) is again the work most similar to the conceptual representation that I presented in this chapter. Appraisal Theory provides a framework for analyzing evaluation and stance in discourse, in context and below the level of the sentence. The three main concepts (systems) in the Appraisal framework correspond to different types of attitudes: *Affect*, which focuses on emotional responses and dispositions, *Judgement*, which is concerned with evaluating human behavior, and *Appreciation*, which is used for evaluating products and processes. For all of these concepts, Appraisal Theory distinguishes between positive and negative variants and the degree of force (intensity). Appraisal Theory also has a system for *Engagement*, which is used to capture hedging, modality, and evidentiality, among other types of subjectivity.

Although the there is an overlap in the types of attitude represented by Appraisal Theory

and the set of attitude types that I present in this chapter, the two representations are not the same. I do not distinguish between affect and the different types of evaluations. Instead, these all fall under the general *sentiment* category in my representation. In addition, the set of attitudes that I propose includes several types of attitude that are not represented at all in the Appraisal framework, such as, agreement and intention. Appraisal Theory also does not include a representation for the target of attitudes, and there is no notion that a single span of text can express more than one type of attitude.

Other text corpora have been developed with annotations of positive and negative sentiments (Yu and Hatzivassiloglou, 2003; Bethard et al., 2004; Kim and Hovy, 2004; Hu and Liu, 2004). In contrast to the below-the-sentence attitude annotations presented in this chapter, the corpora developed by Yu and Hatzivassiloglou (2003), Bethard et al. (2004), and Kim and Hovy (2004), only provide sentence-level annotations. The corpus developed by Hu and Liu (2004) is a bit different from the others. As I do, they annotate targets, specifically products and product features in review data. However they do not then mark the spans of text that express positive and negative sentiments about the targets. Instead, sentiment is annotated as an attribute of the target annotations. It simply captures whether in the sentence there is a positive or negative sentiment toward the target.

7.5 CONCLUSIONS

In this chapter, I extended the original conceptual representation for private states to better model attitudes and their targets. The extension includes a new, more clearly defined set of attitude types and new annotation frames for attitudes and targets. This new representation gives the flexibility needed to associate multiple attitudes and targets with a single private state. Also in the new conceptualization, I redefine intensity explicitly in terms of the new set of attitude types.

I hypothesized that the new scheme for attitudes and targets could be reliably annotated. To test this hypothesis, I conducted two different inter-annotator agreement studies with one other annotator. F-measure agreement for attitude frames was 0.885, and agreement for

targets was on average 0.855. Inter-annotator agreement for attitude types was measured in terms of Cohen's κ over the sets of attitude frames marked by both annotators. Average κ-agreement for attitude type annotations across the two studies was high: 0.80. The results of these studies support the hypothesis that attitudes and targets can be reliably annotated.

I also had hypothesized that defining intensity in terms of the attitude types would lead to more consistent intensity annotations between annotators. Unfortunately, the results of the agreement studies did not provide evidence in support of this hypothesis.

8.0 RECOGNIZING ATTITUDE TYPES: SENTIMENT AND ARGUING

This chapter explores the problem of automatically recognizing *sentiment* and *arguing* attitudes. The attitude annotations in Chapter 7 showed that a large majority of the attitudes in the MPQA Corpus fall into these two general categories. Thus, I focus on recognizing these types of attitude, including their positive and negative variants.

Attitudes can vary greatly in terms of the spans of text on which they are marked. There are attitude annotations that are anchored to single words, as well as attitudes that span entire sentences. This raises the question of what exactly should be classified in the attitude recognition experiments. One possibility is just to classify the attitude of sentences. However, one of the main goals of this dissertation is to develop automatic systems for performing subjectivity analysis below the level of the sentence. With this in mind, I focus my experiments on a new level of fine-grained analysis: classifying the attitude of **attribution levels**. The attribution levels in a sentence are defined based on the direct subjective and speech event (DSSE) expressions in the sentence. I investigate classifying the attitude of attribution levels that are defined based on the manual DSSE annotations, as well as attribution levels defined based on automatically identified DSSEs.

In the attitude classification experiments, I use the clues in the subjectivity lexicon from Chapter 6 as one source of features. When exploring the distribution in sentences of positive and negative clues, I found that many more sentences have mixtures of positive and negative clues than actually have mixtures of positive and negative contextual polarity annotations (see Section 6.3 in Chapter 6). This suggests that disambiguating the contextual polarity of subjectivity clues may help to improve the results for higher-level classification tasks, such as attribution-level sentiment classification. Using the expression-level classifiers from Chapter 6 and a new expression-level subjectivity classifier that I train using the subjective

expressions in the MPQA Corpus, I explore this hypothesis throughout the experiments in this chapter.

This chapter is organized as follows. Sections 8.1 and 8.2, briefly describe the data and the lexicon used in the experiments. The lexicon is the one used in Chapter 6 but with some new information added to the lexicon entries. Section 8.3 describes how the attribution levels and their gold-standard classes are defined. Section 8.4 gives an overview of the expression-level classifiers used to disambiguate clues from the lexicon for the attitude classification experiments. Section 8.5 describes the features used in the attitude classification experiments, which are presented in Section 8.6. The chapter ends with a short overview of related work in Section 8.7 and conclusions in Section 8.8.

8.1 DATASETS

There are essentially two datasets used in this chapter. The **attitude dataset** contains 284 MPQA Corpus (version 2.0) documents with attitude annotations. This dataset is used for 10-fold cross-validation in the attitude classification experiments in Section 8.6. The second dataset is the collection of 494 MPQA Corpus (version 1.2) documents that was used for the contextual polarity experiments in Chapter 6. I call this the **full dataset**. In this chapter, the full dataset is used to train and evaluate the part of the system that automatically identifies units for attitude classification (Section 8.3.1). In addition, the full dataset was used to develop the expression-level classifiers. The expression-level classifiers are used to disambiguate instances of the subjectivity clues from the lexicon, which in turn are used in defining features for attitude classification.

The attitude dataset is actually a subset of the full dataset. Thus, to ensure that the test folds used for cross validation are consistent across the two datasets, the test folds were created as follows. First the 4,499 sentences from the smaller attitude dataset are randomly assigned to the different folds. Then the 5,788 sentences from the remaining documents in the full dataset are randomly assigned to the folds.

8.2 SUBJECTIVITY LEXICON

The subjectivity lexicon that I use for the attitude classification experiments is the same one that I used for the contextual polarity experiments in Chapter 6. Recall that each clue in the lexicon is tagged with two pieces of information, its **reliability class** and its **prior polarity**. A clue's reliability class is either strongly subjective (*strongsubj*) or weakly subjective (*weaksubj*). A clue's prior polarity is *positive*, *negative*, *both*, or *neutral*.

For the experiments in this chapter, I added an additional piece of information to the lexicon for each clue: the clue's **prior arguing polarity**. Prior arguing polarity is intended to capture whether a word out of context seems like it would be used to argue for or against something, or to argue that something is or is not true. A clue's prior arguing polarity may be *positive*, *negative*, or *neutral*. Examples of words with positive arguing polarity are *accuse*, *must*, and *absolutely*. Examples of words with negative arguing polarity are *deny*, *impossible*, and *rather*. Only 2.6% of clues from the lexicon were marked as having a positive arguing polarity, and even fewer, 1.8%, were marked as having a negative arguing polarity.

8.3 UNITS OF CLASSIFICATION

Before turning to the classification experiments, it is first necessary to determine what units will be classified. The text spans of the attitude annotations do not lend an obvious choice for the unit of classification — attitude frames may be anchored to any span of words in a sentence. However, the attitude frames are linked to direct subjective frames, which raises the possibility of trying to classify the attitude of different **attribution levels**.

Each direct subjective or objective speech event annotation in a sentence represents an attribution level. Conceptually, the the text included in an attribution level is the direct subjective or speech event (DSSE) phrase and everything inside the scope of the that phrase. For example, in the sentence below there is an objective speech event frame for the writer of the sentence and two direct subjective frames.

(8.1) [implicit] African observers generally approved of his victory while Western govern-

ments denounced it.

The writer's speech event is implicit, so the attribution level for the writer simply includes all the words in the sentence (the entire sentence being inside the scope of the writer's speech event). The text included in the attribution level represented by the first direct subject frame is the span "generally approved of his victory." The text included in the attribution level represented by the second direct subjective frame is the span "denounced it."

The challenge to working with levels of attribution is identifying the text for each level automatically. There are two parts to this problem. The first is identifying the DSSEs that correspond to the attribution levels. The second part of the problem is defining the full span of text to be included in the attribution level represented by the DSSE. Below I describe the classifier that I use to recognize DSSEs. I then describe the heuristic that I use to identify the text for a given attribution level. The section ends with a description of how I define the gold-standard attitude classes for the attribution levels.

8.3.1 Identifying DSSEs Automatically

The classifier that I use to identify DSSE phrases was developed by Eric Breck. The learning method and features used for this classifier are nearly the same as those used by Breck et al. (2007) for recognizing the combined set of direct subjective frames and expressive subjective elements. The classifier is a sequence-tagging model trained using conditional random fields (Lafferty, McCallum, and Pereira, 2001). It classifies each token as being part of a DSSE or not part of DSSE; any successive string of words tagged as part of a DSSE is considered a DSSE phrase. In addition to features based on words and their contexts, the model includes features based on verbs from Levin's verb classes (Levin, 1993), verbs and adjectives from FrameNet (Framenet, 2002) with frame "experiencer," information about synsets and hypernyms from WordNet 1.6 (Fellbaum, 1998), and reliability class information from the subjectivity lexicon. The Breck et al. tagger for direct subjective frames and expressive subjective elements (Breck, Choi, and Cardie, 2007) uses all these features, as well as features with information about constituent types obtained from a partial parser.

The DSSE classifier was trained and evaluated using 10-fold cross-validation over the

10,287 sentences in the full dataset. The metrics for evaluation are precision and recall. Recall is the percentage of manual, non-implicit DSSE frames identified by the automatic tagger. Precision is the percentage of automatically identified DSSE phrases that are correct. If a DSSE phrase identified by the classifier overlaps with the text span of a DSSE frame annotation, it is considered correct. Using this evaluation, the average recall for the DSSE classifier is 75% and the average precision is 87%.

Breck's DSSE phrase identifier quite correctly does not identify implicit speech events. However, almost all implicit speech events are speech events for the writer of the sentence, which makes them trivial to identify automatically — simply create a DSSE for each sentence and mark it as implicit. Thus, the set of automatically identified DSSEs includes the implicit DSSE for each sentence as well as the DSSEs identified by Breck's classifier.

8.3.2 Defining Levels of Attribution

Unlike the DSSEs, the text in the attribution levels represented by the DSSEs are not marked in the corpus, with the exception of DSSEs that are implicit. For implicit DSSEs (e.g., the DSSE for the writer of the sentence), the text for the attribution level is just the text of the entire sentence. For DSSEs that are not implicit, I use some simple heuristics to define the text for a given attribution level using the DSSE phrase and the dependency parse tree for the sentence.

The way that I define the text for each attribution level is similar to how I defined clauses in Chapter 5 for the intensity classification experiments. Given a DSSE phrase (either from the manual annotations or identified automatically), I first use the parse tree of the sentence to determine which word in the phrase is at the root of the phrase's subtree. The text for the attribution level represented by the DSSE is then all the text in the subtree rooted at that word[1].

Figure 8.1 gives the dependency parse tree and the attribution levels for the sentence *I think people are happy because Chavez has fallen.* The DSSE phrases are in uppercase italics. There is also a DSSE for the writer of the sentence that is not shown because the

[1]For the rare cases where there are two distinct subtrees rooted in a DSSE phrase, I include the text from both subtrees in text for the attribution level.

DSSE phrase is implicit. However, the attribution level for the writer's implicit DSSE is given in the figure. The second attribution level is represented by the DSSE phrase "think." Because "think" is at the root of the tree, the attribution level represented by "think" also includes the entire sentence. The third attribution level is represented by the DSSE phrase "are happy." The word "are" is at the root of the subtree for this phrase. Therefore, the text for the attribution level represented by this DSSE is the part of the sentence contained in the subtree rooted at the word "are."

The above heuristic for identifying the text of attribution levels seems to work well[2] with one glaring exception. In the MPQA Corpus, a number of speech events are marked with the phrase *according to*. However, in a dependency parse tree, "according" is typically a leaf node, even though in most cases the text for the attribution level is the entire sentence. Thus, when a DSSE phrase includes the word "according," I make the the entire sentence the text for the corresponding attribution level.

It is not possible to evaluate the performance of the above heuristic specifically for identifying the text of attribution levels. However, it is possible to evaluate whether the attitude annotations linked to the DSSE frames are encompassed by the text of the corresponding attribution levels. Specifically, for each attitude annotation that is linked to a non-implicit DSSE frame[3], I calculate what percentage of words from the text anchor for the attitude fall within the text of the corresponding attribution level (as defined by the DSSE frame to which the attitude is linked). Table 8.1 shows the results of this evaluation for the 4,243 attitudes linked to non-implicit DSSEs in the attitude dataset. The top row in the table represents the percentage of words from an attitude text span that fall within the text span of the corresponding attribution level. The bottom row of the table gives the number of attitudes that fall within each percentage bin. For example, there are 44 attitudes in which 60% to 79% of the words in those attitudes fall within the text span of the corresponding attribution levels. This evaluation shows that a very large percentage (91%) of attitudes fall entirely within their corresponding attribution levels. The percentage is even higher if the the 1,496 attitudes linked to implicit direct subjective frames are considered.

[2]This is based on a visual inspection, there being no data to use for evaluation.

[3]Attitudes linked to implicit DSSE frames are excluded from this evaluation because by default 100% of the words in the text span of the attitude will be in the text for the attribution level.

Figure 8.1: Attribution levels for the sentence: *I think people are happy because Chavez has fallen*

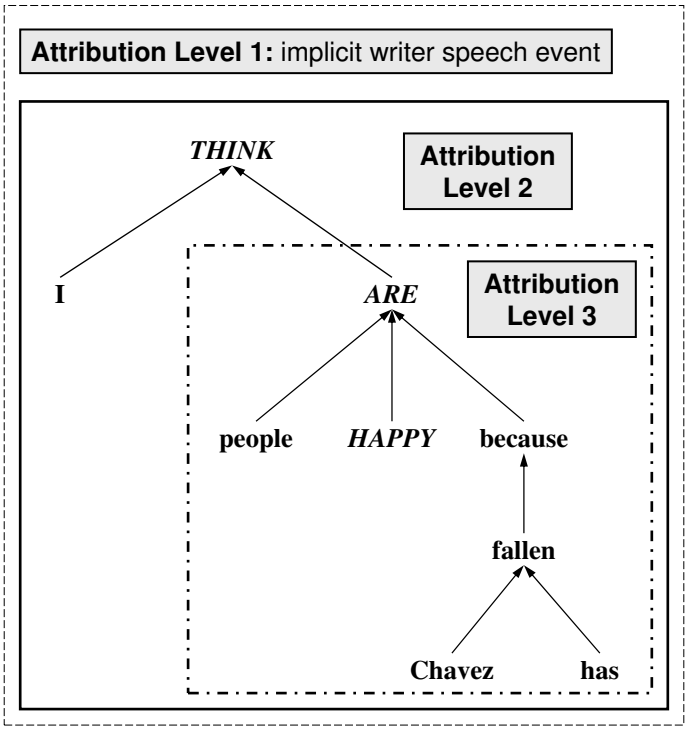


Table 8.1: Counts of attitudes with the given percentage of words contained in the text of the attitudes' corresponding attribution level

percentage of words	0-19%	20-39%	40-59%	60-79%	80-99%	100%
number of attitudes	180	46	60	44	21	3879

There is certainly noise in this method for automatically defining the text spans of the attribution levels. Words undoubtedly are being included in attribution levels where they do not belong. However, with this method over 91% of attitudes fall entirely within their corresponding attribution levels. This inspires confidence that the pertinent information for identifying the different attitudes at least is being included.

8.3.3 Defining the Gold Standard Classes

In the experiments in Section 8.6, I explore classifying the attitude of attribution levels based on both manual DSSE annotations and automatically identified DSSEs. Determining the gold-standard, attitude classes for the attribution levels represented by the manual DSSE frames is straightforward. Each DSSE that is not an objective speech event (recall that DSSEs include both direct subjective frames and objective speech events) will be linked to one or more attitude frames. If the DSSE for an attribution level is linked to an attitude with an *intensity* greater than *low*, then the gold class for that for that attitude type for that attribution level, is true. Otherwise, the gold class for that attitude type and attribution level is false.[4]

Determining the gold-standard attitude classes for the attribution levels based on the automatic DSSEs is only slightly more involved. First the automatic DSSEs are aligned with the manual DSSEs. With one each per sentence, the implicit automatic DSSEs and implicit manual DSSEs are straightforward to align. For the non-implicit DSSEs, a manual

[4]Very occasionally, there is a sentence with more than one implicit DSSE. For these rare cases, I merge all implicit DSSEs into a single implicit DSSE, which inherits all attitudes attached to any implicit DSSE in the sentence.

DSSE is aligned with an automatic DSSE if the two have overlapping text spans. The gold-standard attitude classes for the attribution levels then are defined according to the attitudes linked to the manual DSSEs. Automatic DSSEs that are not aligned to any manual DSSEs will have a gold class that is false for all attitude types.

8.4 EXPRESSION-LEVEL CLASSIFIERS

One hypothesis that I explore in this chapter is that the low-level disambiguation of subjectivity clues is useful for higher-level classification tasks. Specifically, I am interested in investigating whether using expression-level polarity and subjectivity classifiers to disambiguate clue instances will result in improved performance for attitude classification. I use three classifiers in testing this hypothesis. Two of them are expression-level classifiers from Chapter 6, and the third is a new expression-level subjectivity classifier.

The first expression-level classifier is the BoosTexter neutral-polar classifier from Chapter 6 (Section 6.7.1.1) that was trained using all the neutral-polar features. This classifier has an accuracy of 76.5%. It is used to disambiguate clue instances for the *sentiment* recognition experiments.

The second expression-level classifier that I use is the one-step BoosTexter polarity classifier trained using the combined set of neutral-polar and polarity features (see Section 6.7.3 in Chapter 6). This classifier has an accuracy of 74.3%. The output of this classifier is used to disambiguate clue instances for the *positive sentiment* and *negative sentiment* experiments.

The third expression-level classifier is a subjective-expression classifier. This classifier is trained using BoosTexter and the set of neutral-polar features, with the exception of the *polarity modification* features (see Section 6.6.1 in Chapter 6). It has an accuracy of 77.4% (80.1% recall and 80.2% precision) for recognizing subjective clue instances. A baseline classifier that marks every strongly subjective clue instance as subjective has an accuracy of 61.3%.

The output of the subjective-expression classifier is used in two places. First, it is used to determine which clue instances to consider when determining the values of the *clue synset*

features, regardless of the attitude classifier. The *clue synset* features are described in Section 8.5.2. Output from the subjective-expression classifier also is used by the arguing classifiers to restrict the set of clue instances to consider for other features as well. Ideally, I would like to train an an expression-level arguing classifier. However, there are no expression-level arguing annotations in the MPQA Corpus to use in training such a classifier. The hope is that using an expression-level subjectivity classifier to at least weed out objective clue instances will help with arguing classification.

8.5 FEATURES

I use five types of features in the classification experiments in Section 8.6: *bag-of-word* features, *clueset* features, *clue synset* features, *DSSE word* features, and *DSSE wordnet* features. The *bag-of-word* features are straightforward. For a given attribution level, bag-of-words is just the words in the text for that attribution level. I describe each of the remaining types of features below.

8.5.1 Clueset Features

There are four *clueset* features defined for each attitude type that is classified. The value of a *clueset* feature is the number of instances of clues from the clueset that appear within the text of the attribution level. The cluesets are defined based on *reliability class*, which comes from the subjectivity lexicon, and *attitude class*, which comes either from one of the expression-level classifiers (Section 8.4) or from the lexicon.

For sentiment recognition, the *clueset* features are the following:

strongsubj:sentiment-yes
strongsubj:sentiment-no
weaksubj:sentiment-yes
weaksubj:sentiment-no

The reliability class (strongsubj or weaksubj) for a clue instance comes from the clue's entry in the lexicon. Whether a clue instance is sentiment-yes or sentiment-no comes either

from the lexicon or from from the neutral-polar expression classifier. If information from the lexicon is used, clues with a prior polarity of *neutral* are sentiment-no; all other are sentiment-yes. If the neutral-polar expression classifier is used as the source of the sentiment information, then only those clue instances identified as *polar* by the expression-level classifier are categorized as sentiment-yes.

For positive-sentiment classification, the *clueset* features are:

strongsubj:pos-sentiment-yes
strongsubj:pos-sentiment-no
weaksubj:pos-sentiment-yes
weaksubj:pos-sentiment-no

The reliability class for a clue instance as always comes from the clue's entry in the lexicon. Whether a clue instance is pos-sentiment-yes or pos-sentiment-no comes either from the lexicon or from the expression-level polarity classifier. If information from the lexicon is used, then a clue instance is pos-sentiment-yes if it has a *positive* or *both* prior polarity; otherwise, a clue instance is pos-sentiment-no. If the expression-level polarity classifier is used as the source of polarity information, then a clue instance is pos-sentiment-yes only if classified as *positive* or *both* by the expression-level classifier. The *clueset* features for negative-sentiment classification are defined in a similar way.

For arguing recognition, the *clueset* features are:

strongsubj:arguing-yes
strongsubj:arguing-no
weaksubj:arguing-yes
weaksubj:arguing-no

Both the reliability class for a clue instance and whether a clue instance is arguing-yes or arguing-no is taken from the lexicon. Clues listed in the lexicon with a *positive* or *negative* prior arguing polarity are arguing-yes, and all others are arguing-no.

Although there is no expression-level arguing classifier for disambiguating the contextual arguing polarity of each clue instance, a possible way to improve the quality of the *clueset* features for arguing is to constrain the set of clue instances used to only those identified as subjective by the subjective-expression classifier. This is the approach used to disambiguate the clue instances for the various arguing classification experiments.

For positive-arguing classification, the *clueset* features are:

strongsubj:pos-arguing-yes
strongsubj:pos-arguing-no
weaksubj:pos-arguing-yes
weaksubj:pos-arguing-no

The *clueset* features are similar for negative arguing. As with the more general arguing classification, reliability class information is obtained from the lexicon, although the set of clue instances considered is constrained by the subjective-expression classifier as described in the paragraph above. Information about a clue instance's arguing polarity is also obtained from the lexicon, but it is combined with information about negation terms in the surrounding context.

Negation for arguing is not always the same as negation for sentiment. The expression-level polarity classifier looks in the preceding four words for a negation term that is not part of an intensifying phrase. Examples of phrases that intensify rather than negate are *not only* and *nothing if not*. For some types of arguing clues being negated, the negation term will come before, for example, *not true* and *I don't believe*. However, to negate modals such as *should* and *must*, which are often good arguing clues, the negation term follows the clue: *should not* or *must not*.

To incorporate negation when determining the arguing polarity for a clue instance, I do the following. If there is a negation term in the four words preceding the clue instance or in the two words following the instance, and if the negation term is not part of an intensifying phrase, then I assume the instance is being negated. If the instance is being negated and it has a positive arguing polarity in the lexicon, I count the instance as a negative-arguing clue. Similarly, if the instance is being negated and it has a negative arguing polarity in the lexicon, I count the instance as a positive-arguing clue.

8.5.2 Clue Synset Features

A **synsets** is a set of synonymous words and phrases. It is also the basic unit of organization in the WordNet (Fellbaum, 1998) lexical database. Example 8.2 below is a synset from WordNet with its gloss.

(8.2) good, right, ripe – (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes")

The motivation for the *clue synset* features is that there may be useful groupings of clues, beyond those defined in the subjectivity lexicon, that a learning algorithm could exploit for attitude classification. WordNet synsets provide one way of grouping clues.

To define the *clue synset* features, the first step was to extract the synsets for every clue from WordNet 2.0 and add this information to the lexicon for each clue. Every synset in WordNet has a unique identifier; these identifiers are what was added to the lexicon. Then, the *clue synset* feature for a given attribution level is the union of the synsets of all the subjective clue instances that are found in the attribution level. The subjective clue instances are determined based on the output of the subjective-expression classifier.

8.5.3 DSSE Features

The motivation for the DSSE features is that DSSEs, being at the root of the attribution level, might be particularly important when it comes to recognizing attitude type. There are two types of features based on DSSEs: the *DSSE word* features and the *DSSE wordnet* features.

The *DSSE word* feature for an attribution level is just the set of words in the DSSE phrase. If there is no DSSE phrase because the DSSE is implicit, then the value for this features is a special *implicit* token.

There are two *DSSE wordnet* features: DSSE synsets and DSSE hypernyms. The *DSSE synsets* feature is the union of the WordNet synsets for all the words in the DSSE phrase, with the exception of the words in the following stoplist:

is, am, are, be, been, will, had, has, have, having, do, does

Hypernymy is one type of semantic relation defined between synsets in WordNet. The **hypernyms** for a word are thus the synsets that are parents of the synsets to which the word belongs. A hypernym may be the direct parent synset, or a hypernym parent synset further up the tree. Example 8.3 below gives all the hypernyms for the noun synset: *good, goodness (moral excellence or admirableness).*

(8.3)

- morality (concern with the distinction between good and evil or right and wrong; right or good conduct)
- quality (an essential and distinguishing attribute of something or someone)
- attribute (an abstraction belonging to or characteristic of an entity)
- abstraction, abstract entity (a general concept formed by extracting common features from specific examples)
- entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

The *DSSE hypernyms* feature is the union of all the WordNet hypernyms for all the words in the DSSE phrase, with the exception of the words in the above stoplist. If there is no DSSE phrase because the DSSE is implicit, then the value for both the *DSSE synsets* and *DSSE hypernyms* features is a special *implicit* token.

8.6 EXPERIMENTS

The primary goal of the experiments in this section is to test the following two hypotheses. The first is that automatic systems for recognizing sentiment and arguing attitudes can be developed using the features described above in Section 8.5, and that these systems will perform better than baseline systems. The second hypothesis is that disambiguating the polarity and subjectivity of clues instances used in attitude classification will result in improved performance. Because more than one attitude may be expressed at a given attribution level or within a single sentence, all classifiers developed in these experiments are binary classifiers. Each classifier is trained to recognize only one type of attitude. For example, a positive-sentiment classifier distinguishes distinguish between *positive sentiment* and *not-positive sentiment*.

For these experiments, I use BoosTexter and SVM-light. The boosting classifiers are trained using 1,000 rounds of boosting. For the SVM classifiers, I use a radial basis function kernel with a gamma of 0.007. The number of rounds of boosting, the SVM kernel, and the value for gamma as usual were chosen based on experiments on a separate development set.

I also experiment with simple, hand-crafted, rule-based classifiers. There are essentially two rule-based classifiers: RB-cluelex and RB-clueauto. RB-cluelex only uses information about clue instances from the lexicon to make its prediction. RB-clueauto makes predictions using information about clue instances obtained from one of the expression-level classifiers. Which expression-level classification information is used depends on the attitude being classified. When classifying sentiment, RB-clueauto uses information from the expression-level *neutral-polar* classifier. When classifying positive or negative sentiment, RB-clueauto uses information from the expression-level *polarity* classifier. For arguing, positive arguing, or negative arguing, RB-clueauto uses information from the expression-level *subjectivity* classifier, but combines this information with information about clue instances from the subjectivity lexicon. Table 8.2 shows the rule used by each rule-based classifier for each attitude type. For the positive arguing and negative arguing RB-clueauto classifiers, a clue instance is negated if a negation term that is not part of an intensifying phrase is found within a window of four words before or two words after.

Each type of classifier provides a different baseline. The baseline for the rule-based classifiers is RB-cluelex. This the is classifier that uses only prior knowledge from the lexicon to predict the attitude class. It is very simple and straightforward, but also reveals how far basic knowledge about words from the subjectivity lexicon will take us for recognizing sentiment and arguing attitudes. The RB-cluelex classifier also provides the needed point of comparison for evaluating whether the clue instance disambiguation used by RB-clueauto is helpful. For both boosting and SVM, the baseline classifier is the one that uses only the *bag-of-word* features. As mentioned in previous chapters, the set of words that are used in a sentence or clause often provide sufficient information to produce a classifier that performs reasonably well. With the exception of the negative-arguing baselines and the RB-cluelex baseline for positive sentiment, both types of baselines (RB-cluelex and bag-of-words) outperform a most-frequent-class baseline for the various attitude types that are evaluated.

To investigate whether disambiguating the clue instances helps to improve performance for attitude classification, I compare the results for two sets of experiments. In the first set, the values of the *clueset* and *clue synset* features are determined in part based on the output of one of three expression-level classifiers, as described in Section 8.5. In the second set of

Table 8.2: Rules used by rule-based classifiers

Attitude	**RB-cluelex** true if finds:	**RB-clueauto** true if finds:
Sentiment	*strongsubj* clue instance with *positive*, *negative*, or *both* prior polarity	clue instance identified as *polar* by expr-level classifier
Positive Sentiment	*strongsubj* clue instance with *positive* or *both* prior polarity	clue instance identified as *positive or both* by expr-level classifier
Negative Sentiment	*strongsubj* clue instance with *negative* or *both* prior polarity	clue instance identified as *negative* or *both* by expr-level classifier
Arguing	*strongsubj* clue instance with *positive* or *negative* prior arguing polarity	clue instance identified as *subjective* by expr-level classifier and with *positive* or *negative* prior arguing polarity
Positive Arguing	*strongsubj* clue instance with *positive* prior arguing polarity	clue instance identified as *subjective* by expr-level classifier and either a) it has *positive* prior arguing polarity and is not negated in context, or b) it has *negative* prior arguing polarity and it is negated in context
Negative Arguing	*strongsubj* clue instance with *negative* prior arguing polarity	clue instance identified as *subjective* by expr-level classifier and either a) it has *negative* prior arguing polarity and is not negated in context, or b) it has *positive* prior arguing polarity and it is negated in context

experiments, the expression-level classifiers are not used to disambiguate the clue instances, and the values of the *clueset* and *clue synset* features are determined using only information about the clues from the lexicon. If the hypothesis holds that disambiguating the contextual polarity and subjectivity of clue instances is useful for attitude classification, the first set of experiments should achieve better results than the second set.

When determining the features for a given attribution level, unless otherwise noted I exclude all information contained in nested attribution levels. In practice, this affects only the *bag-of-words*, *clueset*, and *clue synset* features, and the clue instances considered for the rule-based classifiers. As an example of what excluding nested information means for these features, consider the attribution levels in Figure 8.1 for the sentence *I think people are happy because Chavez has fallen.* According to the dependency parse tree for the sentence, attribution level 3 is nested inside of attribution level 2, and attribution level 2 is nested inside of attribution level 1. In determining the *bag-of-words*, *clueset*, and *clue synset* features for attribution level 2, all information in attribution level 3 is excluded. This means that none of the words from attribution level 3 will be included in the *bag-of-words* feature for attribution level 2. Similarly, the clue instance "happy" will not be considered when determining the values for the *clueset* and *clue synset* features for attribution level 2. Although excluding nested information in this way often results in little or no information remaining at the outer attribution level for the writer of the sentence, early experiments showed that excluding nested information generally gives better results.

This section begins by reporting results for general sentiment and arguing classification (Section 8.6.1). I then turn to experiments for recognizing the finer-grained attitude categories: positive and negative sentiment (Section 8.6.2) and positive and negative arguing (Section 8.6.3). In Section 8.6.4, I investigate the performance gains that result from disambiguating the clue instances used as features for attitude classification. Finally, in Section 8.6.5, I consider how the classifiers will perform on the attribution levels defined using the automatic DSSES, and in Section 8.6.6 I investigate classifying the attitude of sentences. Table 8.3 gives the distributions of the different attitude types for each of these three units of classification: attribution levels based on manual DSSEs, attribution levels based on automatic DSSEs, and sentences.

Table 8.3: Distribution of attitude types in test Data

Attribution Levels – Manual DSSEs

Total	Sentiment	PosSent	NegSent	Arguing	PosArg	NegArg
9,354	2,430 (26%)	872 (9%)	1,671 (18%)	1,436 (15%)	1,191 (13%)	324 (3%)

Attribution Levels – Automatic DSSEs

Total	Sentiment	PosSent	NegSent	Arguing	PosArg	NegArg
8,720	2,029 (23%)	713 (8%)	1,418 (16%)	1,333 (15%)	1,106 (13%)	303 (3%)

Sentences

Total	Sentiment	PosSent	NegSent	Arguing	PosArg	NegArg
4,499	1,821 (40%)	760 (17%)	1,327 (29%)	1,301 (29%)	760 (17%)	314 (7%)

The experiments that follow are performed using 10-fold cross validation over the 4,499 sentences of the attitude dataset, and the results reported are averages over the 10 folds. Improvements of one experiment over another are evaluated using a two-sided t-test. Unless otherwise noted, $p < 0.05$ is used as the threshold for statistical significance.

8.6.1 Classification Results: Sentiment and Arguing Attitudes

The results for general sentiment and arguing classification at the attribution level are given in Table 8.4. The two rule-based classifiers, RB-clueauto and RB-cluelex, are as described above in Table 8.2. For the boosting and SVM experiments, a variety of classifiers were trained using different feature combinations. The first classifier (the baseline) listed for each algorithm uses just the *bag-of-words* (BAG) features. The second classifier uses the *clueset* and *bag-of-words* features. The third, fourth, and fifth classifiers each build on the second classifier. Classifier three uses the *clue synset* features together with the *clueset* and *BAG* features; classifier four uses the *DSSE words*, *clueset*, and *BAG* features; and classifier five adds the *DSSE wordnet* features to the features from the second classifier. These three classifiers are used to evaluate the performance of the different features. The sixth classifier uses all the features. The last classifier in the list uses all the features except for bag-of-words. For each classifier in Table 8.4, overall accuracy is reported, followed by the recall

(R), precision (P) and F-measure (F) for the attitude type being classified. The F-measures for the the not-sentiment (¬Sent) and not-arguing (¬Arg) classes are also given. The best boosting and SVM results for sentiment and arguing are given in bold.

8.6.1.1 Analysis of Sentiment Classification Results The first observation is that RB-clueauto and the various boosting and SVM classifiers all improve nicely over their respective baselines. RB-clueauto, which uses the output of the neutral-polar expression-level classifier, significantly outperforms ($p < 0.01$) RB-cluelex for all metrics. With the exception of precision for Boosting(2), Boosting(3), SVM(2), and SVM(4), all improvements for the boosting and SVM classifiers over the respective bag-of-words baselines are statistically significant.

The best performing sentiment classifier is the SVM classifier that uses all the features, SVM(6). This classifier achieves an accuracy of 84.9% and a sentiment precision of 80.9%. Improvements over the bag-of-words baseline range from 3.7% for ¬Sent F-measure to 68.3% for sentiment recall. The best performing boosting sentiment classifier is Boosting(7), which uses all the features except for bag-of-words. With the exception of sentiment recall, this classifier outperforms Boosting(6), with a significant improvement in sentiment precision. Also, the best SVM classifier and the best boosting classifier handily outperform ($p < 0.01$) the RB-clueauto classifier with the exception of sentiment recall. Boosting(7) and SVM(6) have sentiment precisions that respectively are 17% and 23.5% higher than that of RB-clueauto.

To evaluate the performance of the *clue synset*, *DSSE words*, and *DSSE wordnet* features, I compare the performance of classifiers (3), (4), and (5) to the performance of classifier (2). For boosting, all three of the classifiers improve over Boosting(2), although only the improvements in sentiment recall for Boosting(4) and Boosting(5) and sentiment F-measure for Boosting(5) are statistically significant. For SVM, the results are similar. The SVM-version of classifiers (3), (4), and (5) all improve over SVM(2), with a number of the improvements being significant, including sentiment F-measure, for all three classifiers. These results show that all three of the different types of features, *clue synset*, *DSSE words*, and *DSSE wordnet*, are useful for sentiment classification. However, the *DSSE wordnet* features are clearly the

Table 8.4: Classification results for sentiment and arguing: Manual DSSEs

	Sentiment				¬Sent	Arguing				¬Arg
Rule-based	ACC	R	P	F	F	ACC	R	P	F	F
(1) RB-cluelex	78.5	53.2	57.9	56.2	85.7	85.4	21.2	57.0	30.9	91.9
(2) RB-clueauto	81.1	57.5	65.5	61.2	87.5	86.1	32.1	58.9	41.6	92.1

	Sentiment				¬Sent	Arguing				¬Arg
Boosting	ACC	R	P	F	F	ACC	R	P	F	F
(1) BAG	79.8	41.8	68.1	51.8	87.2	86.2	34.7	58.8	43.6	92.2
(2) BAG,clueset	81.9	53.5	69.9	60.6	88.2	86.8	38.1	61.0	46.9	92.4
(3) + clue synsets	82.4	54.2	71.3	61.6	88.6	87.0	39.2	61.8	48.0	92.5
(4) + DSSE words	82.7	55.8	71.7	62.7	88.8	87.1	40.5	62.4	49.2	92.6
(5) + DSSE wordnet	83.2	57.1	72.6	63.9	89.1	87.0	41.5	61.2	49.5	92.5
(6) ALL	83.5	**57.7**	73.3	64.6	89.2	87.1	**41.7**	62.0	**49.8**	92.6
(7) - BAG	**84.4**	57.1	**76.8**	**65.5**	**89.9**	**87.4**	40.0	**64.4**	49.3	**92.8**

	Sentiment				¬Sent	Arguing				¬Arg
SVM	ACC	R	P	F	F	ACC	R	P	F	F
(1) BAG	79.3	32.8	72.3	45.1	87.2	86.5	24.9	66.2	36.2	92.5
(2) BAG,clueset	82.1	50.3	72.5	59.4	88.5	87.1	32.1	67.1	43.4	92.7
(3) + clue synsets	83.1	51.8	75.7	61.5	89.2	87.4	33.9	68.8	45.4	92.9
(4) + DSSE words	83.1	52.7	74.9	61.9	89.1	87.4	33.1	69.0	44.7	92.9
(5) + DSSE wordnet	84.8	54.6	80.7	65.2	90.3	**88.0**	34.4	**73.0**	46.7	**93.2**
(6) ALL	**84.9**	**55.2**	**80.9**	**65.7**	**90.4**	87.9	**34.9**	72.4	**47.1**	**93.2**
(7) - BAG	84.7	54.9	79.9	65.1	90.2	87.2	32.0	67.2	43.4	92.8

best performing of three. Classifier (5) performs the best out of the three classifiers, in terms of both sentiment precision and recall, for both boosting and SVM. In fact for SVM, the performance of classifier (5) is only slightly lower than the best classifier, which combines all the features.

8.6.1.2 Analysis of Arguing Classification Results Turning to arguing classification, the most immediate observation is that performance in general is lower than for sentiment classification. The high accuracies are due to the very skewed class distribution (see Table 8.3). F-measures for arguing are all below 50, largely due to low arguing recall.

As with the sentiment classifiers, the various arguing classifiers do outperform their respective baselines, although fewer of the improvements are significant. Arguing recall and F-measure for RB-clueauto is significantly higher than for RB-cluelex. For boosting, the classifiers using the *DSSE words* or *DSSE wordnet* features (classifiers (4), (5), (6) and (7)) also achieve significant improvements in arguing recall and F-measure, and the improvement in precision for Boosting(7) is significant for $p < 0.06$. For SVM, all the classifiers again achieve significantly higher arguing recall and F-measure when compared to the SVM bag-of-words baseline. SVM(5) and SVM(6) in addition achieve significantly higher accuracies and precisions.

The classifiers with the best performance for arguing are the same ones that give the best performance for sentiment. For boosting, this is classifier (7), which uses all the features except for bag-of-words. Boosting(7) improves over the bag-of-words baseline by 15% for arguing recall and 9.5% for arguing precision. For SVM, classifiers (5) and (6) again have the best performance. SVM(5) has a slightly higher precision, and SVM(6) has a slightly higher recall; otherwise there is very little difference between the two in their performance. SVM(6), which combines all the features, achieves a 40% improvement in arguing recall and a 9.4% improvement in arguing precision over the SVM baseline. The *DSSE wordnet* features again seem to have an important role in achieving the best performance, particularly for SVM.

8.6.1.3 Comparison with Upper Bound To get a better understanding for how well the sentiment and arguing classifiers are performing, I compare their results with the upper

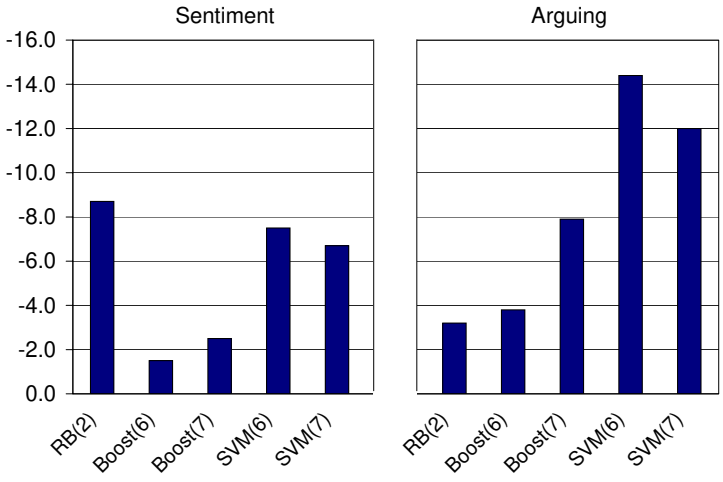
bounds provided by the inter-annotator agreement study. Although the studies in Chapter 7 reported agreement for attitudes, these inter-annotator agreement numbers do not provide the needed comparison. First, the attitude agreement calculated in Chapter 7 considered all attitude types, rather than treating sentiment and arguing attitudes individually as they are in these experiments. Second, agreement was calculated across all matching attitude frames, not across attribution levels represented by DSSEs.

To get a meaningful upper bound requires calculating agreement individually for sentiment and arguing for the set of DSSEs. Although DSSE-level attitudes are easy enough to derive for each annotator from the linked attitude frames, the agreement-level calculated will be inflated. Recall that attitudes were annotated as an additional layer of annotation on top of the existing direct subjective annotations. Agreement for sentiment and arguing attitudes for the subset of DSSEs that are direct subjective frames is κ=0.61 (83%) for sentiment and κ=0.69 (87%) for arguing[5]. Calculating attitude agreement for all DSSEs makes the assumption that annotators would have perfect agreement for identifying DSSEs and distinguishing between direct subjective frames and objective speech event frames, which is obviously not the case. This is the source of the inflated agreement. Considering the set of all DSSEs included in the documents annotated for the agreement studies in Chapter 7, average κ=0.80 (91%) for sentiment and κ=0.76 (93%) for arguing. Although the agreement numbers for DSSE-level attitudes are higher than they would be in reality, as an upper bound, they are still useful.

Comparing the best results from Table 8.4 shows that accuracy is not too much lower than percent agreement, especially considering the inflated agreement. For recognizing sentiment, boosting experiment (7) has an accuracy of 84.4%, and SVM experiment (6) has an accuracy of 84.9%. For arguing, boosting (7) has an accuracy of 87.4%, and SVM (6) has an accuracy of 87.9%. However, when comparing κ for these experiments, the story is different, indicating that there is still quite a bit of room for improvement. For sentiment classification, boosting (7) achieves κ=0.56 and κ for SVM (6) is even lower, 0.42. The kappa results for arguing for these experiments are nearly identical.

[5]These agreement numbers represent an average over the results calculated from the two studies presented in Chapter 7.

Figure 8.2: Decrease in F-measure that results when information from nested attribution levels is included for the given experiments



8.6.1.4 Including Information from Nested Attribution Levels Before turning to the next section, it is interesting to consider exactly how much performance degrades when information from nested attribution levels is included rather than excluded. Figure 8.2 shows the absolute drop in sentiment and arguing F-measure that results when nested information is included during training and evaluation. The results for this experiment are given for several different classifiers. RB(2) is the RB-clueauto classifier. Boost(6) and SVM(6) are the boosting and SVM classifiers that use all the features. Boost(7) and SVM(7) are the classifiers that use all the features except for bag-of-words.

The results in Figure 8.2 show that including information from nested attribution levels degrades the performance of all three types of classifiers: rule-based, boosting, and SVM. This is not unexpected, given the initial experiments that led to the exclusion of nested information. However, some interesting findings are revealed. Boosting seems to be more resilient than SVM to the noise created by the inclusion of nested information. For both sentiment and arguing classification, the performance of the boosting classifiers does not degrade as much as the performance of the SVM classifiers. Comparing the sentiment and arguing classifiers, it is interesting to see that it is always the classifier that performed the

best when nested information was *excluded* that suffers the largest drop in performance when the nested information is *included*. For example, of the boosting classifiers, Boost(7) had the better performance for both sentiment and arguing classification; when nested information is included, of the two classifiers Boost(7) it has the larger drops in F-measure. The same is true for the SVM classifiers, with the best performing classifier SVM(6) also suffering the largest drops in F-measure.

8.6.2 Classification Results: Positive and Negative Sentiment

Table 8.5 gives the positive and negative sentiment classification results for the two rule-based classifiers and a subset of the boosting and SVM classifiers. The boosting and SVM classifiers are the same as those described in the previous section, with the exception of the *clueset* features. These features are determined using the output of the expression-level polarity classifier, as described in Section 8.5.

Many of the observations about the results for classifying positive and negative sentiment are unsurprisingly similar to the results for recognizing the more general sentiment category. Most of the classifiers improve over their respective baselines, and the majority of the improvements are significant. The *DSSE wordnet* features again are extremely helpful: The classifiers using these features consistently achieve the highest precision for both positive and negative sentiment.

What is interesting to analyze are the differences between the respective positive-sentiment and negative-sentiment classifiers. Although RB-clueauto improves over RB-cluelex for both positive-sentiment and negative-sentiment classification, only the improvements for positive sentiment are significant. Still, the RB-clueauto classifier performs better for negative sentiment than for positive sentiment, as measure by sentiment recall, precision, and F-measure. However, when the various features are combined together for the machine learning classifiers, suddenly the picture changes. The positive-sentiment classifiers now achieve levels of performance much higher than the negative-sentiment classifiers. For example, Boosting(7) has a positive-sentiment recall of 54.6% and a precision of 74.1%. These represent improvements of 20% and 36% over the positive-sentiment recall and precision achieved by

Table 8.5: Classification results for positive and negative sentiment: Manual DSSEs

		Pos Sentiment			¬PosSent		Neg Sentiment			¬NegSent
Rule-based	ACC	R	P	F	F	ACC	R	P	F	F
(1) RB-cluelex	89.1	46.7	42.3	44.4	93.9	83.9	45.7	56.2	50.4	90.4
(2) RB-clueauto	91.4	45.6	54.4	49.6	95.3	84.8	48.5	59.2	53.3	90.9

		Pos Sentiment			¬PosSent		Neg Sentiment			¬NegSent
BoosTexter	ACC	R	P	F	F	ACC	R	P	F	F
(1) BAG	92.3	37.7	65.2	47.8	95.9	84.4	31.2	62.8	41.7	91.0
(2) BAG,clueauto	92.3	40.9	63.8	49.8	95.8	85.0	41.6	62.2	49.8	91.2
(5) + DSSE wordnet	93.7	49.4	73.8	59.2	96.6	85.9	43.2	66.0	52.3	91.7
(7) All - BAG	94.0	54.6	74.1	62.9	96.7	86.6	42.2	71.3	53.0	92.2

		Pos Sentiment			¬PosSent		Neg Sentiment			¬NegSent
SVM	ACC	R	P	F	F	ACC	R	P	F	F
(1) BAG	92.0	19.0	79.7	30.6	95.8	84.4	19.9	74.3	31.4	91.2
(2) BAG,clueauto	92.8	35.1	74.1	47.6	96.1	86.1	39.3	69.9	50.3	91.9
(5) + DSSE wordnet	94.5	50.3	85.1	63.2	97.1	86.9	38.4	76.8	51.2	92.4
(6) ALL	94.7	50.8	86.7	64.1	97.1	87.2	39.6	78.4	52.6	92.6

Table 8.6: Classification results for positive and negative arguing: Manual DSSEs

	Pos Arguing				¬PosArg	Neg Arguing				¬NegArg
Rule-based	ACC	R	P	F	F	ACC	R	P	F	F
(1) RB-cluelex	87.3	18.9	49.9	27.4	93.0	96.3	12.3	41.5	19.0	98.1
(2) RB-clueauto	87.8	29.8	53.8	38.4	93.2	96.6	23.6	52.6	32.6	98.2

	Pos Arguing				¬PosArg	Neg Arguing				¬NegArg
Boosting	ACC	R	P	F	F	ACC	R	P	F	F
(1) BAG	87.9	30.1	54.6	38.8	93.3	96.1	21.1	37.9	27.1	98.0
(6) ALL	88.8	37.8	59.3	46.2	93.7	96.5	30.8	49.2	37.9	98.2

	Pos Arguing				¬PosArg	Neg Arguing				¬NegArg
SVM	ACC	R	P	F	F	ACC	R	P	F	F
(1) BAG	88.3	18.8	64.8	29.1	93.6	96.7	07.4	82.3	13.5	98.3
(8) ALL - DSSE wordnet	89.3	31.5	67.6	42.9	94.1	96.9	15.2	78.6	25.5	98.4

the RB-clueauto classifier. For classifying negative sentiment, Boosting(7) achieves a precision of 71.3%, which is not much lower than the precision for positive sentiment. Its recall, however, is quite a bit lower, only 42.2%. This actually represents a 13% *drop* in negative-sentiment recall as compared to RB-clueauto. The results for SVM(6) are very similar. SVM(6) achieves a very high positive-sentiment precision (87%) and a positive-sentiment recall that is 10% higher than the recall for RB-clueauto. For negative sentiment, SVM(6) also achieves a fairly high precision (78.6%), but still a precision lower than that for positive sentiment. Also, it suffers an 18% drop in negative-sentiment recall as compared to RB-clueauto.

8.6.3 Classification Results: Positive and Negative Arguing

Table 8.6 gives the positive and negative arguing results for the rule-based classifiers, the boosting and SVM baselines, and the best performing boosting and SVM classifiers. Interestingly, the best boosting and SVM classifiers are different than for the previous attitude classification experiments. The best boosting classifier is the one that uses all the features,

including bag-of-words. The best SVM classifier is a new classifier, SVM(8), trained using all the features except for the *DSSE wordnet* features.

The results for positive and negative arguing are low. Nevertheless, the various classifiers do achieve significant improvements over their respective baselines. RB-clueauto has a significantly higher recall and F-measure than RB-cluelex for both positive and negative arguing. Boosting(6) achieves significant improvements over the bag-of-words baseline for positive and negative arguing recall, precision and F-measure. SVM(8) also has significantly higher positive arguing recalls and F-measures than SVM baseline.

8.6.4 Benefit of Clue-Instance Disambiguation

There is clearly some benefit to disambiguating the clue instances used in attitude classification. Both RB-cluelex and RB-clueauto use instances of clues from the lexicon to classify attitudes. However, RB-clueauto incorporates information about these clues from different expression-level classifiers, and it outperforms RB-cluelex for all types of attitude classification. Nearly all of the improvements are significant. Still, RB-clueauto is not the best classifier for any of the attitude types, so the question remains: Is there a benefit to disambiguating clues for the best attitude classifiers, when all the different features are combined and working together?

To answer this question, I retrained the best classifiers for each attitude type using versions of the *clueset* and *clue synset* features that do not rely on output from the expression-level classifiers (see Section 8.5). Then, I compared the results of these new classifiers to the results of the original classifiers that incorporated the clue-instance disambiguation.

Table 8.7 shows the changes in recall and precision for sentiment, positive sentiment, and negative sentiment classification for this experiment using the Boosting(7) and SVM(6) classifiers. With the exception of positive-sentiment classification, not performing clue-instance disambiguation does result in lower recalls and precisions for sentiment classification. Although the drop in performance is not large, the results do show promise. Perhaps with better clue-instance disambiguation, significant improvements in attribution-level sentiment classification could be achieved.

Table 8.7: Changes in recall and precision for best sentiment classifiers without clue disambiguation

	Sentiment		Pos Sent		Neg Sent	
Classifier	R	P	R	P	R	P
Boosting(7)	-0.6	-0.7	+1.3	-0.1	-1.8	-1.8
SVM(6)	-1.5	-0.5	+0.2	+0.4	-1.3	-0.4

Table 8.8: Changes in recall and precision for best arguing classifiers without clue disambiguation

	Arguing		Pos Arguing		Neg Arguing	
Classifier	R	P	R	P	R	P
Boosting(7/6)	-1.1	-2.8	-2.9	**-3.5**	-2.7	-2.6
SVM(6/8)	-1.0	+1.7	**-3.7**	-1.8	-5.3	+2.5

The results for this experiment for arguing, positive arguing, and negative arguing classification are shown in Table 8.8. For arguing classification, the classifiers used in this experiment are Boosting(7) and SVM(6). Recall that it was the Boosting(6) and SVM(8) classifiers that achieved the best results for positive and negative arguing classification, so they are the classifiers used for those attitude types in this experiment.

As low are the results are for the various types of arguing classification, without clue-instance disambiguation, results are for the most part even lower. Because of the high variance in the performance of the classification folds, none of the differences in recall and precision are statistically significant for $p < 0.05$. However, the two values in bold in the table are significant for $p < 0.1$. As with the clue-instance disambiguation results for sentiment classification, the evidence that disambiguation is useful for arguing classification is not strong. However, the results do suggest that clue-instance disambiguate has the potential to be useful for higher-level arguing classification.

Table 8.9: Classification results for sentiment and arguing: Automatic DSSEs

Trained Manual		Sentiment			¬Sent		Arguing			¬Arg
Tested Automatic	ACC	R	P	F	F	ACC	R	P	F	F
Boosting(7)	83.9	57.1	68.6	62.3	89.9	86.9	41.7	60.4	49.3	92.5
SVM(6)	85.0	56.0	73.1	63.4	90.5	87.8	34.3	70.6	46.1	93.1

Trained Automatic		Sentiment			¬Sent		Arguing			¬Arg
Tested Automatic	ACC	R	P	F	F	ACC	R	P	F	F
Boosting(7)	83.8	52.0	70.9	60.0	89.9	86.8	38.0	60.6	46.7	92.4
SVM(6)	84.7	49.7	76.3	60.2	90.5	87.8	34.3	70.6	46.1	93.1

Trained Manual		Sentiment			¬Sent		Arguing			¬Arg
Tested Manual	ACC	R	P	F	F	ACC	R	P	F	F
Boosting(7)	84.4	57.1	76.8	65.5	89.9	87.4	40.0	64.4	49.3	92.8
SVM(6)	84.9	55.2	80.9	65.7	90.4	87.9	34.9	72.4	47.1	93.2

8.6.5 Results with Automatic DSSEs

All the results in the previous sections were for classifying the attitude of attribution levels determined using the manual DSSEs (*manual attribution levels*). However, manual DSSEs will not be available in practice, and it is important to know how well the attitude classifiers will perform on attribution levels that are based on the automatic DSSEs (*automatic attribution levels*). To investigate the answer to this question, I conducted two additional sets of experiments with the Boosting(7) and SVM(6) sentiment and arguing classifiers. First, I took the existing classifiers that were trained on the manual attribution levels and tested them on the automatic attribution levels. For the second set of experiments, I retrained the classifiers on the automatic attribution levels and tested them on the automatic levels.

Table 8.9 shows the results of these experiments. For comparison, the results for the Boosting(7) and SVM(6) classifiers both trained and tested on the manual attribution levels (from Section 8.6.1) are given at the bottom of the table. The original sentiment classifiers trained on manual attribution levels and tested on the automatic attribution levels achieve precisions of 68.6 and 73.1 for Boosting(7) and SVM(6), respectively. These precisions

are about 10% lower than the precisions for these same classifiers tested on the manual attribution levels (results at the bottom of the table). Precision for the arguing classifiers trained on manual attribution levels and tested on automatic levels are also lower, but the differences are smaller. Although at first glance it seems like there is very little difference in recall for training on manual and testing on automatic for these classifiers, these two recall values are not actually comparable. The manual DSSEs and the automatic DSSEs are strongly overlapping, but they are not identical sets.

In the middle of Table 8.9 are the results for training and testing the sentiment and arguing classifiers on the automatic attribution levels. This actually results in an increase in precision over the classifiers trained on the manual attribution levels. However, the increase in precision comes at the cost of a lower recall, resulting in lower F-measures overall.

8.6.6 Sentence-level Attitude Classification

Although the focus of this chapter is sentiment and arguing classification at the attribution level, many NLP applications are interested in the attitude of sentences. This raises the question of how well features for classifying the attitude of attribution levels will perform for classifying the attitudes of sentences.

The results for sentence-level sentiment classification, including positive-sentiment and negative-sentiment classification, are given in Table 8.10. For each type of sentiment classification, the table shows results for four different classifiers, with the highest results given in bold. The first two classifiers are the RB-cluelex and RB-clueauto classifiers used previously. The remaining two classifiers are SVM classifiers. Classifier (3) uses *bag-of-words*, *clueset*, *clue synset*, and *DSSE wordnet* features. The *DSSE wordnet* features for a sentence are simply the union of the *DSSE wordnet* features for every automatic DSSE found in the sentence. Classifier (4) uses the same features as classifier (3), but without using the output of the expression-level classifiers to disambiguate the clue instances used in determining the *clueset* and *clue synset* features. Results for SVM *bag-of-words* classifiers are not given because they performed lower than the respective RB-clueauto classifiers.

For determining whether or not a sentence is expressing a sentiment, it turns out that

Table 8.10: Results for sentence-level sentiment classification

Sentiment	ACC	R	P	F	¬F
(1) RB-cluelex	75.2	68.5	69.6	69.1	79.3
(2) RB-clueauto	79.2†	**70.4**	76.3†	73.3†	83.0†
(3) SVM all features	**80.9**	68.6	**81.3**	**74.4**	**84.8**
(4) (3) with no disambig	79.6	67.0	79.5	72.7	83.8

Pos Sentiment	ACC	R	P	F	¬F
(1) RB-cluelex	82.4	**55.2**	48.1	51.4	89.2
(2) RB-clueauto	86.5†	53.5	61.1†	**57.0**†	92.0†
(3) SVM all features	**88.4**	43.0†	**78.7**	55.6†	**93.3**
(4) (3) with no disambig	87.6	37.6	77.3	50.6	92.9

Neg Sentiment	ACC	R	P	F	¬F
(1) RB-cluelex	79.4	59.8	66.8	63.1	85.7
(2) RB-clueauto	81.5	**62.3**	71.6†	**66.6**	87.2
(3) SVM all features	**82.7**	54.6	**80.8**	65.2	**88.5**
(4) (3) with no disambig	82.6	54.8	79.9	65.0	88.5

the RB-clueauto is difficult to beat. Although other machine learning classifiers with various combinations of features were tried, only SVM produced classifiers that outperformed RB-clueauto. Classifier (3) achieves significantly higher sentiment precision and ¬sentiment F-measure than RB-clueauto, and the improvement in accuracy is significant for $p < 0.06$. For positive sentiment and negative sentiment classification, classifier (3) also achieves higher accuracy, precision, ¬sentiment F-measure than RB-clueauto. These improvements are significant for positive sentiment; for negative sentiment, the improvements in recall and ¬sentiment F-measure are significant.

Table 8.11 reports results for sentence-level arguing classification, with the highest result for each metric in bold. The first two classifiers are RB-cluelex and RB-clueauto. The third classifier is an SVM classifier trained using *bag-of-words* and *clueset* features. This is the best of the sentence-level arguing classifiers. The last classifier is the same as classifier (3), but without clue disambiguation for the *clueset* features. The best SVM classifier (row 3) performs significantly better than RB-clueauto for all metrics with the exception of arguing precision.

Table 8.11: Results for sentence-level arguing classification

Arguing	ACC	R	P	F	¬F
(1) RB-cluelex	75.4	26.9	69.4	38.7	84.6
(2) RB-clueauto	78.0†	40.0†	71.5	51.3†	85.8†
(3) SVM BAG,clueset	**81.1**	**52.4**	**74.9**	**61.7**	**87.5**
(4) (3) with no disambig	80.5	49.5	74.6	59.5	87.2

Finally, I again consider whether disambiguating clue instances is helpful for higher-level attitude classification, this time at the sentence level. The classifiers that do use the clue instance disambiguation provided by the expression-level classifiers (classifiers 2 and 3 in Tables 8.10 and 8.11) perform better than the corresponding classifiers that do not make use of clue disambiguation (classifiers 1 and 4) for almost all metrics. Where the improvements are statistically significant is indicated by a dagger (†) in the tables.

8.7 RELATED WORK

Attribution levels have not specifically been the focus of research in subjective-objective or attitude-type classification. However, research by Choi et al. (2006) and Breck et al. (2007) on identifying direct subjective expressions effectively results in the first step to identifying a large subset of the subjective attribution levels. Choi et al. (2006) work on jointly identifying direct subjective frames and their sources. Breck et al. (2007) focuses on identifying both direct subjective expressions and expressive subjective elements. One attribution level that I include in my classification experiments, but that Choi et al. (2006) and Breck et al. (2007) do not include, is the attribution level for the speaker of the sentence. Work by Breck and Cardie (2004) recovers the nested structure of attribution levels but does not try to make any classification with respect to sentiment or other attitude.

The most closely related research to the work in this chapter on sentiment classification, is the work on classifying positive and negative sentiments at the sentence level. Yu and

Hatzivassiloglou (2003), Nasukawa and Yi (2003), Kim and Hovy (2004), Hu and Liu (2004), Kudo and Matsumoto (2004), Popescu and Etzioni (2005), and Gamon et al. (2005) have all worked on identifying sentences expressing positive and negative sentiments. Yu and Hatzivassiloglou, Nasukawa and Yi, Kim and Hovy, and Hu and Liu classify sentiment sentences by aggregating information about words from a lexicon. My experiments also rely on information from a lexicon, but I combine that information with other features that are new to sentiment classification. Kudo and Matsumoto, Popescu and Etzioni, and Gamon et al. apply different machine learning approaches to sentence-level sentiment classification. Kudo and Matsumoto use a boosting algorithm that takes into account the structure of a sentences. Popescu and Etzioni use an approach called relaxation labelling to jointly find the sentiment of words, product features and sentence. Gamon et al. train a Naive Bayes classifier using Expectation Maximization (EM) to bootstrap from a small set of labelled data. In contrast to the approach that I take to attitude classification, none of the above sentence-level approaches allow for a sentence to be tagged as having both a positive or negative sentiment.

WordNet has primarily been used in sentiment classification to identify lists of positive and negative words (e.g., Kamps and Marx (2002), Hu and Liu (2004), Kim and Hovy (2004), Esuli and Sebastiani (2005), and Andreevskaia and Bergler(2006)), which are often then used as input features for sentiment classification. In their work on identifying opinion expressions in context, Choi et al. (2006) and Breck et al. (2007) use WordNet in a novel way for subjectivity analysis—using hypernyms as input features for classification. It is their use of WordNet in this way that inspired some of the features that I use in this chapter for sentiment and arguing classification.

Liu et al. (2003) and Gordon et al. (2003) have also worked on classification tasks involving attitude categories, however the categories they investigate are very different. Liu et al. seek to identify when a sentence is expressing one of the basic emotion categories proposed by Eckman (1992). The goal of the work of Gordon et al. is to classify expressions of commonsense psychology. To the best of my knowledge, this chapter presents the first research in automatically recognizing arguing attitudes at the sentence-level or below.

8.8 CONCLUSIONS

In this chapter, I explored the automatic recognition of sentiment and arguing attitudes at the attribution level and at the sentence level. I experimented with three different types of algorithms for recognizing the attitude of attribution levels: rule-based, boosting, and SVM. For all of these, I was able to develop classifiers that performed significantly better than the baseline classifiers. In addition, the best boosting and SVM classifiers outperformed the best rule-based classifier, which relied on information from the expression-level classifiers in making its classifications.

As part of my experiments, I investigated the performance of several new types of features. The *clue synset* features capture information from WordNet about what synsets the clues from the subjectivity lexicon belong to. Other features represent information about the direct subjective or speech event phrase for the given attribution level. The best performing of the new features actually did incorporate both types of information, representing the words in the direct subjective or speech event phrase for an attribution level using WordNet synsets and hypernyms.

I hypothesized that expression-level subjectivity and sentiment classifiers could be used to disambiguate clue instances being used as features for higher-level classification tasks, and that this would give improved results for the higher-level classification. For rule-based classification where only the clues from the lexicon are used, disambiguating clues definitely is beneficial. However, for the best classifiers that use a wide variety of features, the increases that result from clue-instance disambiguation are fairly small and rarely statistically significant. Nonetheless, there are improvements. This suggests that as expression-level classifiers are improved the question of whether the output of these classifiers can be used to disambiguate features for higher-level classifiers should be revisited.

9.0 QUESTION ANSWERING AND FINE-GRAINED SUBJECTIVITY ANALYSIS

Question Answering (QA) is one of the NLP applications that I was considering when motivating the need for fine-grained subjectivity analysis in the opening chapter of this dissertation. Although it is beyond the scope of this dissertation to empirically evaluate whether the automatic systems I developed can be used to the benefit of an actual QA system, it is possible to investigate whether fine-grained subjectivity analysis has the *potential* to improve QA. Specifically, in this chapter I explore the interaction between different private state and attitude annotations and the question answers marked in the Opinion Question Answering (OpQA) Corpus.

The OpQA Corpus is a subset of 94 documents from the MPQA Corpus, annotated by Stoyanov et al. (2004; 2005) with answers to fact and opinion questions. All of the OpQA documents are also part of the set of documents with attitude annotations. Table 9.1 lists the questions annotated in the OpQA Corpus. In total, there are 125 answers to the fact questions and 414 answers to the opinion questions.

Stoyanov et al. (2004; 2005) describes two ways in which subjectivity information might be used to improve opinion QA: (1) by assisting in **answer selection** and (2) by helping with **answer ranking**. The idea with answer selection is that the QA system already has retrieved a set of potential answers to a question, and it now must select from those answers the ones that it thinks are best. The hypothesis is that opinion questions will be answered more often in text segments classified as subjective, and that fact questions will be answered more often in text segments classified as objective. The idea with answer ranking is that subjectivity information can be used to help the QA system from the beginning to retrieve the correct answers.

Table 9.1: Fact and opinion questions in the OPQA Corpus

Fact Questions	
What is the Kyoto Protocol about?	
When was the Kyoto Protocol adopted?	
Who is the president of the Kiko Network?	
What is the Kiko Network?	
What is the murder rate in the United States?	
What country issues an annual report on human rights in the United States?	
Who is Andrew Welsdan?	
When did Hugo Chavez become President?	
Which governmental institutions in Venezuela were dissolved by the leaders of the 2002 coup?	
Who is Vice-President of Venezuela?	
Where did Mugabe vote in the 2002 presidential election?	
At which primary school had Mugabe been expected to vote in the 2002 presidential election?	
How long has Mugabe headed his country?	
Who was expecting Mugabe at Mhofu School for the 2002 election?	

Opinion Questions	
Does the president of the Kiko Network approve of the US action concerning the Kyoto Protocol?	Sentiment
Are the Japanese unanimous in their opinion of Bush's position on the Kyoto Protocol?	Sentiment
How is Bush's decision not to ratify the Kyoto Protocol looked upon by Japan and other US allies?	Sentiment
How to European Union countries feel about the US opposition to the Kyoto protocol?	Sentiment
How do the Chinese regard the human rights record of the United States?	Sentiment
What factors influence the way in which the US regards the human rights records of other nations?	
Is the US Annual Human Rights Report received with universal approval around the world?	Sentiment
Did anything surprising happen when Hugo Chavez regained power in Venezuela after he was removed by a coup?	
Did most Venezuelans support the 2002 coup?	Sentiment
How did ordinary Venezuelans feel about the 2002 coup and subsequent events?	Sentiment
Did America support the Venezuelan foreign policy followed by Chavez?	Sentiment
What was the American and British reaction to the re-election of Mugabe?	Sentiment
What is the basis for the European Union and US critical attitude and adversarial action toward Mugabe?	
What did South Africa want Mugabe to do after the 2002 election?	Sentiment
What is Mugabe's opinion about the West's attitude and actions toward the 2002 Zimbabwe election?	Sentiment

For the experiments in this chapter, I replicate and expand on the answer selection experiments conducted by Stoyanov and his colleagues. To investigate whether information about the subjectivity or objectivity of answers could be used to improve answer selection, Stoyanov and colleagues calculated the following conditional probabilities:

(1) Prob(text with answer a is subjective | a answers an opinion question)
(2) Prob(text with answer a is objective | a answers an opinion question)
(3) Prob(text with answer a is subjective | a answers a fact question)
(4) Prob(text with answer a is objective | a answers a fact question)

If taking subjectivity information into account helps with answer selection, then probabilities 1 and 4 will be high, and probabilities 2 and 3 will be low.

9.1 TEXT GRANULARITY AND ANSWER SELECTION

In this dissertation, I have experimented with various levels of fine-grained subjectivity analysis, from individual words to entire sentences. Yet how important is fine-grained analysis to answer selection for QA? To answer this question, I calculate and compare the answer selection probabilities for three different levels of subjective text span granularity. For the finest level of granularity, I use the original private state expression-level annotations. If an answer overlaps with a direct subjective annotation or an expressive subjective element, then it is in a subjective text span. The answer probabilities using this first definition for subjective text spans are given at the top of Table 9.2. To define subjective text spans that are less fine grained than expressions but still more fine grained than sentences, I use the attitude annotations. If an answer overlaps with an attitude annotation, then it is in a subjective text span. The answer probabilities using this definition for subjective text spans are given in the middle of Table 9.2. For the coarsest level of granularity, I let subjective text spans correspond to subjective sentences. An answer is in a subjective text span if it is in a subjective sentence, where a sentence is considered to be subjective if it contains at least one direct subjective annotation. The answer probabilities using the last definition of subjective text span are given at the bottom of Table 9.2.

Table 9.2: Overlap of answers to fact and opinion questions with subjectivity annotations of different levels of granularity

	Question Type			
expression	fact		opinion	
objective	108	(86.4%)	44	(10.6%)
subjective	17	(13.6%)	370	(90.4%)

	Question Type			
attitude	fact		opinion	
objective	88	(70.4%)	51	(12.3%)
subjective	37	(29.6%)	363	(87.7%)

	Question Type			
sentence	fact		opinion	
objective	30	(24.0%)	13	(3.1%)
subjective	95	(76.0%)	401	(96.9%)

Looking at Table 9.2, there are two main observations. First, the probability of an answer to an opinion question being in a subjective sentence is very high, 96.9%. This is helpful for selecting answers to opinion questions, but not very helpful for selecting answers to fact questions: 76% of the answers to fact questions are also in subjective sentences. However, as the level of subjective text span granularity becomes more refined, the answer probabilities for the fact questions improve, while still maintaining high answer probabilities for the opinion questions.

9.2 INTENSITY AND ANSWER SELECTION

The next question I consider is whether answers to opinion questions are more likely to be found in text spans with higher intensities. To answer this question, I calculate the following answer probabilities: given an answer to a fact/opinion question, what is the probability of the answer overlapping with an attitude of a particular intensity. If an answer does

Table 9.3: Overlap of answers with attitude annotations of differing intensities

	Question Type			
intensity	fact		opinion	
neutral	88	(70.4%)	51	(12.3%)
low	7	(13.6%)	12	(2.9%)
low-medium	9	(5.6%)	24	(5.8%)
medium	9	(7.2%)	107	(25.9%)
medium-high	7	(7.2%)	84	(20.3%)
high	5	(4.0%)	136	(30.4%)

not overlap with any attitude annotation, it's intensity is neutral. Table 9.3 gives these probabilities. The results in the table show that indeed answers to opinion questions are more likely to be in subjective text spans with higher intensities. On the other hand, if an answer to a fact question overlaps with an attitude, then that attitude is most likely to have a low intensity.

9.3 SENTIMENT AND ANSWER SELECTION

The last question I consider is whether answers to sentiment questions are more likely to be found in spans of text where a sentiment is being expressed as compared to spans of text that are just generally subjective. To answer this question, I calculate three different sets of conditional probabilities. First, given an answer to a sentiment question/other question, I calculate the probability that an answer overlaps with a subjective annotation that has a positive, negative, or both polarity. Next, given an answer to a particular type of question, I calculate the probability of that answer overlapping with a sentiment attitude annotation. For the last probability, given an answer to a sentiment question/other question, I calculate the probability of the answer overlapping with any attitude annotation. If an answer overlaps with an attitude annotation of any type, I consider it to be in a subjective text span; otherwise, it is in an objective text span. The sentiment questions in the OpQA corpus are

Table 9.4: Overlap of answers to sentiment and non-sentiment questions with polar expressions, sentiment attitude spans, and subjective text spans

	Question Type			
	¬sentiment		sentiment	
neutral	132	(81.5%)	77	(20.4%)
polar	30	(18.5%)	300	(79.6%)

	Question Type			
	¬sentiment		sentiment	
¬sentiment	117	(72.2%)	106	(28.1%)
sentiment	45	(27.8%)	271	(71.9%)

	Question Type			
	¬sentiment		sentiment	
objective	103	(63.6%)	55	(14.6%)
subjective	59	(36.4%)	322	(85.4%)

indicated above in Table 9.1.

For this experiment, I expected that an answer to a sentiment question would have a higher probability of overlapping with a polar expression or with a sentiment attitude than being in a subjective text span. However, the opposite is true. The results are given in Table 9.4. Given that an answer is an answer to a sentiment question, there is 85.4% chance that the answer overlaps with a subjective text span, and only a 71.9% chance that the answer overlaps with a sentiment attitude annotation.

9.4 DISCUSSION

Based on the above experiments, it does seem that finer-grained subjectivity analysis, both in terms of the text spans that are identified and the distinctions in intensity and attitude, has the potential to help with QA. However, rather than helping with opinion QA, the answer selection experiments seem to suggest that fine-grained subjectivity analysis might

be the most helpful for weeding out opinions and sentiments when they are not the kind of answers being sought.

Does this result then mean that for opinion QA, there is no need to go to the extra effort of fine-grained analysis to help, for example, with answering questions that specifically are targeting sentiments? To answer this question, I conduct one more set of experiment comparing the following probabilities: the probability of an answer to a sentiment question being in (1) any sentence, (2) a subjective sentence, and (3) a sentence containing a sentiment attitude. There is an 11.7% chance of any given sentence containing an answer to a sentiment questions. The chance of a subjective sentence containing an answer to a sentiment question is a bit higher, 13.9%. As I had hoped, sentiment sentences have the highest chance of containing an answer to a sentiment questions: 17.1%. This result supports the idea that focusing in on different types of attitudes from the beginning of the QA answer retrieval process may be one way to help with opinion QA.

10.0 RESEARCH IN SUBJECTIVITY AND SENTIMENT ANALYSIS

In the earlier chapters, I compared and contrasted the research that was most closely related to the specific research contributions in the various parts of this dissertation. In this chapter, I give a more general overview of research in automatic subjectivity and sentiment analysis in text.

Research in automatic subjectivity and sentiment analysis ranges from work on learning *a priori* knowledge about the words and phrases associated with subjectivity and sentiment, to applications trying to glean the general mood or sentiment of large populations (e.g., communities of bloggers) from documents on the web. In the sections below, I review the research in some of the main areas of automatic subjectivity and sentiment analysis.

10.1 IDENTIFYING *A PRIORI* SUBJECTIVE INFORMATION ABOUT WORDS AND PHRASES

A number of researchers have worked on identifying words and phrases that are associated with subjective language. Wiebe (2000) and Baroni and Vegnaduzzo (2004) identify subjective adjectives, and Wiebe et al. (2004) identifying subjective verbs and n-grams. Riloff et al. (2003) use two bootstrapping approaches to identify sets of subjective nouns, and Riloff and Wiebe (2003) identify extraction patterns that are correlated with expressions of subjectivity. Kobayashi et al. (2004) identify domain-dependent sets of subjective expressions. Kim and Hovy (2005) use small seed sets and WordNet to identify sets of subjective adjectives and verbs, and Esuli and Sebastiani (2006) proposes a method for identifying both the subjectivity and prior polarity of a word also using WordNet. Wiebe and Mihalcea (2006)

tackle a more fine-grained task, automatically identifying whether a particular word *sense* is subjective.

Hatzivassiloglou and McKeown (1997) were the first to address the problem of acquiring the prior polarity (semantic orientation) of words. Since then this has become a fairly active line of research in the sentiment community with various techniques being proposed for identifying prior polarity. Turney and Littman (2003) and Gamon and Aue (2005) use statistical measures of word association. Kamps and Marx (2002), Hu and Liu (2004), Takamura et al. (2005), Esuli and Sebastiani (2005), and Andreevskaia and Bergler (2006) propose various methods for learning prior polarity from WordNet. Popescu and Etzioni (2005) use an unsupervised classification technique for jointly learning the prior polarity and contextual polarity of words and the polarity of sentences. Kanayama and Nasukawa (2006) use an unsupervised method for acquiring domain dependent polarity lexicons for Japanese. Takamura et al. (2005) proposes a model for learning the prior polarity of phrases.

10.2 IDENTIFYING SUBJECTIVE LANGUAGE AND ITS ASSOCIATED PROPERTIES IN CONTEXT

Research on automatically identifying subjective language and its properties in context encompasses a wide variety of tasks. There is work on subjective and objective sentence classification (Wiebe, Bruce, and O'Hara, 1999; Riloff and Wiebe, 2003; Yu and Hatzivassiloglou, 2003; Wiebe and Riloff, 2005), recognizing expressions of opinions in context (Choi, Breck, and Cardie, 2006; Breck, Choi, and Cardie, 2007), and classifying the intensity of sentences and clauses (Wilson, Wiebe, and Hwa, 2006). Other researchers focus on recognizing the sentiment of phrases or sentences (Morinaga et al., 2002; Yu and Hatzivassiloglou, 2003; Nasukawa and Yi, 2003; Kim and Hovy, 2004; Hu and Liu, 2004; Kudo and Matsumoto, 2004; Hurst and Nigam, 2004; Popescu and Etzioni, 2005; Gamon et al., 2005; Kaji and Kitsuregawa, 2006; Wilson, Wiebe, and Hoffmann, 2005). Bethard et al. (2004), Kim and Hovy (2004; 2006), and Choi et al. (2005; 2006) all work on identifying the source of opinions. Breck and Cardie (2004) identify nested levels of attribution (e.g., that it is according

to China that the U.S. believes something).

10.3 EXPLOITING AUTOMATIC SUBJECTIVITY ANALYSIS IN APPLICATIONS

Many different NLP applications have been proposed or developed that make use of automatic subjectivity and sentiment analysis, including direction-based text retrieval (Hearst, 1992), recognizing inflammatory messages (Spertus, 1997), tracking sentiment timelines in on-line discussions (Tong, 2001), extracting investor sentiment from stock message boards (Das and Chen, 2001), distinguishing editorials from news articles (Wiebe, Wilson, and Bell, 2001; Yu and Hatzivassiloglou, 2003), automatic expressive text-to-speech synthesis (Alm, Roth, and Sproat, 2005), information extraction (Riloff, Wiebe, and Phillips, 2005), question answering (Yu and Hatzivassiloglou, 2003; Stoyanov, Cardie, and Wiebe, 2005), multi-document summarization (Seki et al., 2005), analyzing public comments for pro and con responses (Kwon, Shulman, and Hovy, 2006), determining support and opposition in congressional debates (Thomas, Pang, and Lee, 2006) and recognizing document perspective (Lin et al., 2006; Lin and Hauptmann, 2006).

In the past few years, two applications in particular have been the focus of a great deal of research. The first is mining and summarizing opinions from product reviews (e.g., (Morinaga et al., 2002; Nasukawa and Yi, 2003; Yi et al., 2003; Hu and Liu, 2004; Popescu and Etzioni, 2005; Yi and Niblack, 2005; Carenini, Ng, and Pauls, 2006; Hu and Liu, 2006)). The second is product and movie review classification (e.g., (Turney, 2002; Pang, Lee, and Vaithyanathan, 2002; Morinaga et al., 2002; Dave, Lawrence, and Pennock, 2003; Nasukawa and Yi, 2003; Beineke, Hastie, and Vaithyanathan, 2004; Mullen and Collier, 2004; Pang and Lee, 2005; Whitelaw, Garg, and Argamon, 2005; Kennedy and Inkpen, 2005; Koppel and Schler, 2005; Ng, Dasgupta, and Arifin, 2006; Cui, Mittal, and Datar, 2006)).

11.0 CONCLUSIONS AND DIRECTIONS FOR FUTURE WORK

This dissertation investigated the manual and automatic identification of different types of fine-grained subjectivity in a large corpus of news documents from the world press. For the manual identification, annotators were trained to identify expressions of private states, as well as the component parts of private states: sources, attitudes, and targets. Annotators were also trained to mark properties of private states, such as their intensity.

The automatic identification focused on three types of fine-grained subjectivity analysis: recognizing the intensity of clauses and sentences, recognizing the contextual polarity of words and phrases, and recognizing the attribution levels where sentiment and arguing attitudes are being expressed. A supervised-learning paradigm was used to develop automatic systems for performing each of these types of fine-grained subjectivity analysis.

11.1 SUMMARY OF RESULTS AND CONTRIBUTIONS

11.1.1 The MPQA Corpus

A large part of the work in this dissertation was directed at the production of the MPQA Corpus. Two versions of the corpus have been released by our research group to the larger research community, with one more release planned with the new attitude and target annotations. In addition to supporting our own research and that of our collaborators, in the past few years, the MPQA Corpus has begun to be used by other groups in their published research (Kim and Hovy, 2005; Kim and Hovy, 2006; Seki et al., 2005; Eguchi and Lavrenko, 2006).

11.1.2 Conceptualization and Annotation of Private States

The results of an inter-annotator agreement study conducted as part of this dissertation show that annotators can be trained to reliably annotate expressions of private states as represented in the conceptualization of private states (Wiebe, 2002; Wiebe, Wilson, and Cardie, 2005). This dissertation then extends the conceptualization to better represent the attitudes and targets of private states. The results of subsequent annotation studies confirm that the extensions too can be reliably annotated.

The annotation studies also reveal that intensity can be difficult to annotate reliably. Even when intensity is defined explicitly in terms of attitude type and annotators are given guidelines for annotating different levels of intensity, inter-annotator agreement does not really improve.

11.1.3 Automatic Systems for Fine-Grained Subjectivity Analysis

The experiments in this dissertation show that automatic systems can be developed for performing fine-grained subjectivity analysis. Using different supervised machine learning algorithms, classifiers were trained to identify the intensity of clauses and sentences, the contextual polarity of words, and whether attribution levels are expressing sentiments or arguing attitudes. All of these automatic system outperform the baseline systems developed for the same tasks.

11.1.4 Features for Fine-Grained Subjectivity Analysis

Fine-grained subjectivity analysis requires a wide variety of features to achieve the best results. With the exception of the SVM classifiers trained for recognizing contextual polarity and a few of the classifiers for recognizing arguing attitudes, the classifiers using the greatest number and widest variety of features consistently achieved the best performance for all the different types of fine-grained subjectivity analysis investigated in this dissertation. Features combining lexical information with information about syntactic structure consistently helped to improve performance for both recognizing intensity and recognizing contextual polarity.

11.1.5 Importance of Recognizing Neutral Expressions

The experiments in recognizing contextual polarity show that being able to identify when positive and negative words are being used in neutral contexts is a very important part of the problem. A study of manual contextual polarity annotations shows that positive and negative words are used in neutral contexts much more often than they are used in expressions of the opposite polarity. Furthermore, the noise of neutral instances greatly impairs the ability of features to distinguish between positive and negative contextual polarity.

11.1.6 Disambiguating Subjectivity Clues for Higher-level Classification

An exploration of the distribution of positive and negative clues in sentences suggested that disambiguating the contextual polarity of subjectivity clues would help to improve the results for higher-level sentiment classification tasks. For simple classifiers that rely primarily on information about the clues, this observation is confirmed, but when the clues are used in the presence of other strong features the benefit of disambiguating the clues is diluted and no longer significant.

11.2 DIRECTIONS FOR FUTURE WORK

Fine-grained subjectivity analysis is challenging. Although the work in this dissertation has made strides both in the development of resources for fine-grained subjectivity analysis and in extending our understanding of the types of linguistic information that are useful for automatically recognizing the intensity, contextual polarity, and attitudes of private states, these problems are far from solved. The possibilities of where to go next with this research, run in a number of directions.

11.2.1 Moving to New Domains

The automatic systems for fine-grained subjectivity analysis presented in this dissertation were developed on news data. Yet subjectivity and sentiment analysis is important in many other domains, from weblogs to product reviews to political speeches to consumer call center data. This raises the important question of how well the systems for recognizing intensity, contextual polarity, and attitudes will perform on data other than the news.

One way of approaching this question is by considering how domain dependent or independent are the various features. Obviously the most domain dependent features are the *bag-of-words* features. However, a few of the experiments in this work suggest that *bag-of-words* features may not always be required. For example, one of the contextual polarity experiments investigated whether the word token of a clue instance was needed given the other polarity features. The results of this experiment showed that the word token was not needed to determine polarity if the system already knew that a word was expressing a sentiment. For sentiment and arguing classification, experiments were also conducted without *bag-of-words* features, again with very promising results. To confirm whether the classifiers that do not use *bag-of-words* features are more robust to changes in domain than those that do, future work should include experiments to compare the performance of these classifiers across different domains. Also, future work in developing any new systems for fine-grained sentiment analysis should seek to exclude these most domain-dependant features.

Compared to the *bag-of-words* features, the majority of other features investigated in this research can be expected to be much more domain independent. This is because most other features correspond in some way to *sets* of words or clues. For example, there is one feature for each of the different sets of syntactic clues used for intensity recognition. The polarity modification features capture whether or not a clue instance is modifying/being modified by another clue instance and the prior polarity (polarity set) of that instance. For attitude recognition, the *DSSE wordnet* features use WordNet synsets and hypernym sets. Using features that correspond to sets rather than individual words or clues from the lexicon allows for greater generalizability, which should be good for applying the systems to new domains.

Although many of the features investigated in this research should be fairly generalizable, these features still rely on the clues represented in the subjectivity lexicon or perhaps learned from in-domain data. This raises a couple challenges when it comes to applying the systems developed in this work to new domains. First, although many clues of subjective language and their prior polarities are domain independent, others are not. Related to this is the challenge created by the huge variety in subjective language. New domains mean new subjective terminology that will not be present in the lexicon and therefore unknown to the system. This will certainly result in a degradation of system performance, particularly recall, when applying the system to new domains.

One possible way to overcome the problems of new and domain dependent subjective language would be to create a customized subjectivity lexicon for each new domain. Kanayama and Nasukawa (Kanayama and Nasukawa, 2006), for example, propose an unsupervised method for acquiring domain-dependent polarity lexicons. Given a new domain-specific lexicon, it would be straightforward to substitute the new lexicon for the one currently currently in use by the automatic classifiers. In theory, this should lead to improved performance for the automatic systems on the new domain without the need for retraining. It will be interesting to actually test this in the future to see if the hypothesis holds.

11.2.2 Increasing Knowledge of Subjective Language

To continue to improve automatic fine-grained subjectivity analysis, one of the most important challenges will be to increase our knowledge of how private states are expressed. Working at such a low level means that for a given unit of classification, whether it is a phrase or a clause, there is relatively little information to go on. There is a much greater chance that the system may not have the necessary information to identify the subjectivity in a small unit, than for a larger unit, such as a paragraph or document. The need for more knowledge is clear when looking back over the experiments in this dissertation: Recall is often a problem.

What obviously comes to mind in terms of increasing our knowledge of subjective language is extending the coverage and prior knowledge represented in subjectivity lexicons.

While this is important, especially in terms of understanding how subjective language differs from one domain to another, it is only a small part of the subjective knowledge needed to interpret subjectivity in context, as evidenced by the experiments in this work. More information is needed about the interaction between clues of subjective language and about the types of things that can influence the subjectivity and polarity of words and phrases in context. Also, more information is needed about how best to generalize lexical information for fine-grained subjectivity analysis. Various organizations and sets of words and subjectivity clues were used in the experiments throughout this dissertation. For the most part, these groupings were determined based on human knowledge and intuition. Might there be better ways to organized the clues of subjective language to give even better performance and to allow even greater generalizability? If so, would it be possible to learn these better organizations automatically, perhaps through methods of automatic clustering?

11.3 BEYOND CLASSIFICATION

With the exception of SVM regression used for intensity recognition, all the machine learning methods used in this dissertation were straightforward, supervised classification algorithms. These algorithms represented a variety of different types of learning and provided the needed basis for evaluating features, but there are certainly other approaches that might be considered for future work. For intensity recognition, it would be good to investigate more sophisticated techniques for ordinal regression, such as the large-margin methods proposed by Herbrich et al. (Herbrich, Graepel, and Obermayer, 2000). For recognizing contextual polarity, one possibility that would be interesting to explore is clustering. Hierarchical clustering exploiting distributional similarities has been used successfully for word-sense disambiguation (Lee, 1997), for example, which has similarities to determining the polarity of a word in context. For recognizing the attitude of attribution levels, the precision for recognizing sentiments (including positive/negative) is high enough that bootstrapping approaches to building up classifiers, such as those proposed by Wiebe and Riloff (Wiebe and Riloff, 2005) should be considered in future work.

11.3.1 Extrinsic Evaluation of Automatic Systems

One of the primary motivations for this research is the need for fine-grained subjectivity analysis in applications such as question question answering, information extraction, and summarization. Although the automatic systems developed for intensity, contextual polarity, sentiment and arguing recognition were evaluated *intrinsically*, they have yet to be evaluated *extrinsically*[1]. How much will these systems help in end applications?

As part of an extrinsic evaluation, it again will be important to evaluate the contribution of the various features. The majority of features that were proposed were found to be at least a little helpful for the various tasks. However, the cost in terms of generating the different features is not the same. Features such as the the count of clues in a sentence or the presence of a negation term are very quick to compute. On the other hand, the new syntactic features for intensity recognition, some of the modify features for polarity recognition, and the DSSE features from the attitude recognition experiments all require a dependency parse of the sentence. Compared to underlying NLP technologies such as part-of-speech tagging, parsing is still fairly slow. In addition to creating a problem for using these features in any real-time applications, it also raises the very basic question of whether generating these features will prove worthwhile. In other words, how much of a performance improvement is needed in the underlying fine-grained subjectivity analysis for an improvement to be seen in the end application, and if the improvement provided by a given feature is very small, is it worthwhile to use that feature considering the cost of generating it?

Although it is impossible to answer the above questions without carrying out an actual evaluation, results of the experiments in this work can at least provide intuition into what will translate into noticeable improvements for an end application. Consider, for example sentence-level intensity recognition. For this task, SVM achieved a mean-squared error of 0.75 using all the features. When the syntactic clues are excluded, MSE increases to 0.806, and when only bag-of-words are used, MSE is 0.962. An MSE of 0.75 means that, on average, the predictions of the intensity recognizier are less than one away from the the actual

[1]Earlier versions of the sentence-level sentiment and arguing classifiers were evaluated in the context of a straightforward QA system (Somasundaran et al., 2007), with results suggesting that identifying sentiment and arguing can help with opinion QA.

intensity (on average 0.866 away). There is not not that much difference when the syntactic clues are excluded (on average 0.898 away). However, if just the bag-of-words are used, intensity predictions approach being a full intensity class off (0.981 away on average). Thus, while excluding the syntactic clues for intensity recognition may not result in a noticeable decrease in performance for an end application, using none of the subjectivity clues will likely negatively affect an end application.

The results of the contextual polarity experiments suggest that it is unlikely that excluding individual features (with the possible exception of negation) will result in a noticeable difference for an end application. Even when all the neutral-polar features were excluded for the one-step polarity classifiers, the decreases in performance are probably small enough that they would have little impact on an end application, even though some of the decreases were statistically significant. However, performance for positive and negative recall is much lower if just the word token is used, and including prior polarity only helps with negative recall (positive recall remains low). For example, the accuracy differs by less than three percentage points from the BoosTexter word-token one-step polarity classifier[2] to the all-feature one-step classifier (71.7% to 74.3%), whereas positive recall differs by 6.5 points. This difference in recall likely is enough to affect the performance of an end application.

For attribution-level sentiment and arguing recognition, fewer features were evaluated. The results of those experiments suggest that excluding the *DSSE wordnet* features would be enough to create a difference for an end application. Although absolute improvements in accuracy were small (1–2 percentage points), including the *DSSE wordnet* features typically produced improvements in recall ranging from 5–15 points for sentiment recognition (including positive/negative) and improvements in precision ranging from 3.5–10 points. These differences will quite likely translate into noticeable differences for an end application.

When incorporating the systems presented in this work into an end application, it will be important to carefully consider the trade off between the performance of the individual features and the cost of generating the features. Although there are a couple features that the experiments revealed to be particularly strong and useful (*negation* and *DSSE wordnet* features) that should be included if at all possible (the *DSSE wordnet* features require a

[2]This experiment was conducted as part of this research but not reported in Chapter 6.

parse of the sentence and access to WordNet), depending on the task, a subset of the other features will likely be sufficient. In determining which features to use, it is the recall and precision of the subjective class(s) that should be the determining factor, not accuracy.

11.4 BEYOND INTENSITY, CONTEXTUAL POLARITY, AND ATTITUDES

This dissertation investigated the automatic recognition of the intensity, polarity, and attitudes of private states, but there are other problems of fine-grained subjectivity analysis that will certainly arise as this area of research continues to grow. Given a new task in subjectivity analysis, an important question to consider is what this research tells us about the types of features or algorithms that might be recommended for that task.

Whether or not features should be included to capture syntactic relationships between subjectivity clues will likely depend on the granularity of the task. If the unit of analysis is words, phrases or even clauses, syntactic and structural information seems to be important. For example, the difference in mean-squared error caused by the exclusion of the syntactic features for intensity recognition increases as the clause level being classified becomes more fine grained. Also, the features capturing syntactic relationships between clues were consistently useful for recognizing the contextual polarity of phrases. On the other hand, for sentence-level sentiment and arguing recognition, a number of features were tried that represent interesting relationships between different kinds of subjective knowledge. None of these ended up being reported because they were not found to be at all helpful for this higher level of subjectivity analysis.

Also important for subjectivity analysis are features that represent sets of clues grouped according to prior knowledge about their subjective properties. Such features from this work include the features representing sets of strongly subjective and weakly subjective clues (contextual polarity recognition), features representing sets of clues organized by their intensity (intensity recognition), features representing sets of clues based on their prior polarity (contextual polarity recognition), etc. All of these features proved useful. However, how

subjectivity clues should be organized into sets for defining features should depend on the task. The same organization of clues for one task (i.e., an organization based on intensity), will not necessarily be the best one for other tasks.

Two other features that should be used depending on the task are negation and features capturing information about the DSSE phrase. If the new task involves polarity, even if it is not sentiment polarity, negation features should be incorporated. If the domain for analysis is the news, features representing the DSSE phrase were shown to be very important for the this data, and this can be expected to be true regardless of the type of subjectivity analysis being considered.

Although the intention of the experiments in this dissertation was not to draw conclusions about which learning algorithms are more or less suited to different types of fine-grained subjectivity analysis, some tentative observations can be made about what seemed to work best. BoosTexter performed relatively well for all the different tasks: It always gave better results than Ripper and it had the highest results (along with TiMBL) for recognizing contextual polarity. The contextual polarity experiments incorporated the most diverse types of features out of the different tasks, including features capturing complex relationships between subjectivity clues. Looking at the results of the ablation experiments for these features, the argument can be made that BoosTexter seemed to do the best job of making use of all the different features. Thus, if the new task involves the need to capture complex relationships between clues of subjective language, BoosTexter would likely be a good choice.

SVM-light did not perform as well as BoosTexter (or TiMBL) for recognizing contextual polarity[3] However, for intensity recognition, it achieved the best performance in terms of mean-squared error, and it also had the highest performance for sentiment and arguing recognition. For these tasks, the features that were used were either bag-of-words or features representing sets of clues or words. In other words, there were no features such as were used for recognizing contextual polarity that represented complex relationships between subjectivity clues. If this is also true about the features being considered for a new task in subjectivity analysis, SVM-light would likely be a good choice.

[3]It is possible that this is a result of not choosing the right parameters.

BIBLIOGRAPHY

References

Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 347–354, Vancouver, Canada.

Andreevskaia, Alia and Sabine Bergler. 2006. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from WordNet gloasses. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.

Asher, Nicholas. 1986. Belief in discourse representation theory. *Journal of Philosophical Logic*, 15:127–189.

Bai, Xue, Rema Padman, and Edoardo Airoldi. 2005. On learning parsimonious models for extracting consumer opinions. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 3*.

Baker, C., C. Fillmore, and J. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL-98)*, pages 86–90, Montreal, Canada.

Ballmer, T. and W. Brennenstuhl. 1981. *Speech Act Classification: A Study in the Lexical Analysis of English Speech Activity Verbs*. Springer-Verlag, Berlin; New York.

Banfield, Ann. 1982. *Unspeakable Sentences*. Routledge and Kegan Paul, Boston.

Baroni, M. and S. Vegnaduzzo. 2004. Identifying subjective adjectives through web-based mutual information. In Ernst Buchberger, editor, *Proceedings of KONVENS-04, 7th Konferenz zur Verarbeitung Natürlicher Sprache (German Conference on Natural Language Processing)*, pages 17–24.

Beineke, Philip, Trevor Hastie, and Shivakumar Vaithyanathan. 2004. The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 263–270, Barcelona, Spain.

Bethard, Steven, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *Working Notes — Exploring Attitude and Affect in Text: Theories and Applications (AAAI Spring Symposium Series)*.

Breck, Eric and Claire Cardie. 2004. Playing the telephone game: Determining the hierarchical structure of perspective and speech expressions. In *Proceedings of the Twentieth International Conference on Computational Linguistics (COLING 2004)*, pages 120–126, Geneva, Switzerland.

Breck, Eric, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, Hyderabad, India.

Bruce, Rebecca and Janyce Wiebe. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2):187–205.

Cardie, Claire, Janyce Wiebe, Theresa Wilson, and Diane Litman. 2003. Combining low-level and summary representations of opinions for multi-perspective question answering. In *Working Notes of the AAAI Spring Symposium in New Directions in Question Answering*.

Carenini, Giuseppe, Raymond Ng, and Adam Pauls. 2006. Multi-document summarization of evaluative text. In *11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*.

Chatman, Seymour. 1978. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca, New York.

Choi, Yejin, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 431–439, Sydney, Australia.

Choi, Yejin, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 355–362, Vancouver, Canada.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–20.

Cohen, W. 1996. Learning trees and rules with set-valued features. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 709–717, Portland, Oregon, August. American Association for Artificial Intelligence, Cambridge: MIT Press.

Cohn, Dorrit. 1978. *Transparent Minds: Narrative Modes for Representing Consciousness in Fiction*. Princeton University Press, Princeton, NJ.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 16–23, Madrid, Spain.

Cui, Hang, Vibhu O. Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06*.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 168–175, Philadelphia, Pennsylvania.

Daelemans, Walter, Véronique Hoste, Fien De Meulder, and Bart Naudts. 2003a. Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, pages 84–95.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003b. TiMBL: Tilburg memory Based Learner, version 5.0 Reference Guide. ILK Technical Report 03-10, Induction of Linguistic Knowledge Research Group, Tilburg University. Available at http://http://ilk.uvt.nl/downloads/pub/papers/ilk0310.pdf.

Das, S. R. and M. Y. Chen. 2001. Yahoo! for Amazon: Opinion extraction from small talk on the web. In *Proceedings of the 8th Asia Pacific Finance Association Annual Conference (APFA)*, Bangkok, Thailand.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*, Budapest, Hungary. Available at http://www2003.org.

de Rivera, Joseph. 1977. *A Structural Theory of Emotions*. New York: International Universities Press.

Doležel, Lubomir. 1973. *Narrative Modes in Czech Literature*. University of Toronto Press, Toronto, Canada.

Eguchi, Koji and Victor Lavrenko. 2006. Sentiment retrieval using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 345–354, Sydney, Australia.

Ekman, Paul. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3/4):169–200.

Esuli, Andrea and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM-05)*, pages 617–624, Bremen, Germany.

Esuli, Andrea and Fabrizio Sebastiani. 2006. Determining term subjectivity and term orientation for opinion mining. In *Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 193–200, Trento, IT.

Fauconnier, Gilles. 1985. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge, Massachusetts: MIT Press.

Fellbaum, Christiane, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge.

Fodor, Janet Dean. 1979. *The Linguistic Description of Opaque Contexts*. Outstanding dissertations in linguistics 13. Garland, New York & London.

Framenet. 2002. *http://www.icsi.berkeley.edu/∼framenet/*.

Gamon, M., A. Aue, S. Corston-Oliver, and E. Ringger. 2005. Pulse: Mining customer opinions from free text. In *Intelligent Data Analysis*.

Gamon, Michael. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 611–617, Geneva, Switzerland.

Gamon, Michael and Anthony Aue. 2005. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Ann Arbor, Michigan.

Gordon, Andrew, Abe Kazemzadeh, Anish Nair, and Milena Petrova. 2003. Recognizing expressions of commonsense psychology in English text. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pages 208–215, Sapporo, Japan.

Grefenstette, G., Y. Qu, J.G. Shanahan, and D.A. Evans. 2004. Coupling niche browsers and affect analysis for an opinion mining application. In *Proceedings of the Conference Recherche d'Information Assistee par Ordinateur (RIAO-2004)*.

Halliday, M.A.K. 1985/1994. *An Introduction to Functional Grammar*. London: Edward Arnold.

Hatzivassiloglou, Vasileios and Kathy McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 174–181, Madrid, Spain.

Hearst, Marti. 1992. Direction-based text interpretation as an information access refinement. In Paul Jacobs, editor, *Text-Based Intelligent Systems*. Lawrence Erlbaum.

Heise, David R. 1965. Semantic differential profiles for 1000 most frequent English words. *Psychological Monographs*, 79(601).

Herbrich, Ralf, Thore Graepel, and Klaus Obermayer. 2000. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132.

Hoste, Véronique. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, Language Technology Group, University of Antwerp.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2004 (KDD 2004)*, pages 168–177, Seattle, Washington.

Hu, Minqing and Bing Liu. 2006. Opinion extraction and summarization on the web. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-2006), Nectar Paper Track*, Boston, MA.

Hummel, R.A. and S.W. Zucker. 1983. On the foundations of relaxation labeling processes. *IEEE Transations on Pattern Analysis and Machine Intelligence (PAMI)*, 5(3):167–187.

Hurst, Matthew and Kamal Nigam. 2004. Retrieving topical sentiments from online document collections. In *Proceedings of SPIE, Document Recognition and Retrieval XI*, number 5296 in Proceedings of SPIE, pages 27–34.

Hwa, Rebecca and Adam Lopez. 2004. On converting constituent parses to depedency parses. Technical Report TR-04-118, University of Pittsburgh.

Joachims, T. 1999. Making large-scale SVM learning practical. In B. Scholkopf, C. Burgess, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, Cambridge, MA. MIT-Press.

Johnson-Laird, P.N. and Keith Oatley. 1989. The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, 3(2):81–123.

Kaji, Nobuhiro and Masaru Kitsuregawa. 2006. Automatic construction of polarity-tagged corpus from HTML documents. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 452–459, Sydney, Australia.

Kamps, Jaap and Maarten Marx. 2002. Words with attitude. In *Proceedings of the First International WordNet Conference*, pages 332–341, Mysore, India.

Kanayama, Hiroshi and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 355–363, Sydney, Australia.

Kennedy, Alistair and Diana Inkpen. 2005. Sentiment classification of movie and product reviews using contextual valence shifters. In *Proceedings of FINEXIN-05: Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations*.

Kennedy, Alistair and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the Twentieth International Conference on Computational Linguistics (COLING 2004)*, pages 1267–1373, Geneva, Switzerland.

Kim, Soo-Min and Eduard Hovy. 2005. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005)*, pages 61–66, Jeju Island, KR.

Kim, Soo-Min and Eduard Hovy. 2006. Identifying and analyzing judgment opinions. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 200–207, New York, New York.

Kobayashi, Nozomi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04)*.

Koppel, Moshe and Jonathan Schler. 2005. The importance of neutral examples for learning sentiment. In *Proceedings of FINEXIN-05: Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations*.

Koppel, Moshe and Jonathan Schler. 2006. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology, 2nd Edition*. Sage Publications, Thousand Oaks, California.

Kudo, Taku and Yuji Matsumoto. 2004. A boosting algorithm for classification of semi-structured text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 301–308, Barcelona, Spain.

Kuroda, S.-Y. 1973. Where epistemology, style and grammar meet: A case study from the Japanese. In P. Kiparsky and S. Anderson, editors, *A Festschrift for Morris Halle*. Holt, Rinehart & Winston, New York, NY, pages 377–391.

Kuroda, S.-Y. 1976. Reflections on the foundations of narrative theory—from a linguistic point of view. In T.A. van Dijk, editor, *Pragmatics of Language and Literature*. North-Holland, Amsterdam, pages 107–140.

Kwon, Namhee, Stuart W. Shulman, and Eduard Hovy. 2006. Multidimensional text analysis for eRulemaking. In *Proceedings of the 7th Annual International Conference on Digital Government Research (dg.o 2006)*, pages 157–166.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA.

Lee, Lillian. 1997. *Similarity Based Approaches to Natural Language Processing*. Ph.D. thesis, Harvard University.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL-98)*, pages 768–773, Montreal, Canada.

Lin, Wei-Hao and Alexander Hauptmann. 2006. Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1057–1064, Sydney, Australia.

Lin, Wei-Hao, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the*

10th Conference on Computational Natural Language Learning (CoNLL-2006), pages 109–116, New York, New York.

Liu, Hugo, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI-2003)*, pages 125–132, Miami, Florida.

Martin, J.R. 1992. *English Text: System and Structure*. Philadelphia/Amsterdam: John Benjamins.

Martin, J.R. 2000. Beyond exchange: APPRAISAL systems in English. In Susan Hunston and Geoff Thompson, editors, *Evaluation in Text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press, pages 142–175.

Maybury, Mark T., editor. 2004. *New Directions in Question Answering*. Menlo Park: American Association for Artificial Intelligence.

Morinaga, Satoshi, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. 2002. Mining product reputations on the web. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pages 341–349, Edmonton, Canada.

Mullen, Tony and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 412–418, Barcelona, Spain.

Nasukawa, Tetsuya and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003)*, pages 70–77, Sanibel Island, Florida.

Ng, Vincent, Sajib Dasgupta, and S. M. Niaz Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 611–618, Sydney, Australia.

Ortony, Andrew, Gerald L. Clore, and Mark A. Foss. 1987. The referential structure of the affective lexicon. *Cognitive Science*, 11:341–364.

Osgood, Charles E., G.J. Suci, and P.H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana: University of Illinois Press.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 115–124, Ann Arbor, Michigan.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 79–86, Philadelphia, Pennsylvania.

Polanyi, Livia and Annie Zaenen. 2004. Contextual valence shifters. In *Working Notes — Exploring Attitude and Affect in Text: Theories and Applications (AAAI Spring Symposium Series)*.

Popescu, Ana-Maria and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 339–346, Vancouver, Canada.

Quirk, Randolph, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, New York.

Rapaport, William. 1986. Logical foundations for belief representation. *Cognitive Science*, 10:371–422.

Riloff, Ellen and Rosie Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-1999)*, pages 474–479, Orlando, Florida.

Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 105–112, Sapporo, Japan.

Riloff, Ellen, Janyce Wiebe, and William Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proc. 20th National Conference on Artificial Intelligence (AAAI-2005)*, pages 1106–1111, Pittsburgh, PA.

Riloff, Ellen, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pages 25–32, Edmonton, Canada.

Russell, J.A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Schapire, Robert E. and Yoram Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

Seki, Yohei, Koji Eguchi, Noriko Kando, and Masaki Aono. 2005. Multi-document summarization with subjectivity analysis at DUC 2005. In *Proceedings of the 2005 Document Understanding Conference (DUC-2005)*.

Somasundaran, Swapna, Theresa Wilson, Janyce Wiebe, and Veselin Stoyanov. 2007. QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *International Conference on Weblogs and Social Media*.

Spertus, Ellen. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of the Eighth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-97)*, pages 1058–1065, Providence, Rhode Island.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Stoyanov, Veselin, Claire Cardie, Diane Litman, and Janyce Wiebe. 2004. Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus. In *Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Stoyanov, Veselin, Claire Cardie, and Janyce Wiebe. 2005. Multi-perspective question answering using the opqa corpus. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 923–930, Vancouver, Canada.

Suzuki, Yasuhiro, Hiroya Takamura, and Manabu Okumura. 2006. Application of semi-supervised learning to evaluative expression classification. In *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006)*, pages 502–513, Mexico City, Mexico.

Takamura, Hiroya, Takashi Inui, and Manabu Okumura. 2005. Extracting emotional polarity of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan.

Thelen, M. and E. Riloff. 2002. A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 214–221, Philadelphia, Pennsylvania.

Thomas, Matt, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the Conference on*

Empirical Methods in Natural Language Processing (EMNLP-2006), pages 327–335, Sydney, Australia.

Tong, Richard. 2001. An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the SIGIR Workshop on Operational Text Classification*, pages 1–6, New Orleans, Louisianna.

Turney, Peter. 2002. Thumbs 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 (ACL-02)*, pages 417–424, Philadelphia, Pennsylvania.

Turney, Peter and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Uspensky, Boris. 1973. *A Poetics of Composition*. University of California Press, Berkeley, CA.

Watson, D. and A. Tellegen. 1985. Toward a consensual structure of mood. *Psychological Bulletin*, 98:219–235.

White, P.R.R. 2002. Appraisal: The language of attitudinal evaluation and intersubjective stance. In Verschueren, Ostman, blommaert, and Bulcaen, editors, *The Handbook of Pragmatics*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pages 1–27.

Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM-05)*, pages 625–631.

Wiebe, J. 2002. Instructions for annotating opinions in newspaper articles. Department of Computer Science Technical Report TR-02-101, University of Pittsburgh.

Wiebe, Janyce. 1990. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text*. Ph.D. thesis, State University of New York at Buffalo.

Wiebe, Janyce. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Wiebe, Janyce. 2000. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 735–740, Austin, Texas.

Wiebe, Janyce, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. 2003. Recognizing and organizing opinions expressed in the world press. In *Working Notes of the AAAI Spring Symposium in New Directions in Question Answering*, pages 12–19, Palo Alto, California.

Wiebe, Janyce, Rebecca Bruce, and Thomas O'Hara. 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)*, pages 246–253, College Park, Maryland.

Wiebe, Janyce, Kenneth McKeever, and Rebecca Bruce. 1998. Mapping collocational properties into machine learning features. In *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-6)*, pages 225–233, Montreal, Canada.

Wiebe, Janyce and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of COLING-ACL 2006*.

Wiebe, Janyce and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, pages 486–497, Mexico City, Mexico.

Wiebe, Janyce, Theresa Wilson, and Matthew Bell. 2001. Identifying collocations for recognizing opinions. In *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, pages 24–31, Toulouse, France.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39(2/3):164–210.

Wilks, Yorick and Janusz Bien. 1983. Beliefs, points of view and multiple environments. *Cognitive Science*, 7:95–119.

Wilson, Theresa and Janyce Wiebe. 2003. Annotating opinions in the world press. In *Proceedings of the 4th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*, pages 13–22, Sapporo, Japan.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 347–354, Vancouver, Canada.

Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004)*, pages 761–766, San Jose, California.

Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99.

Yi, Jeonghee, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003)*, pages 427–434, Melbourne, Florida.

Yi, Jeonghee and Wayne Niblack. 2005. Sentiment mining in WebFountain. In *Proceedings the 21st International Conference on Data Engineering (ICDE-05)*, pages 1073–1083, Tokyo, Japan.

Yu, Hong and Vasileios Hatzivassiloglou. 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-2003)*, pages 129–136, Sapporo, Japan.