AMI Consortium

http://www.amiproject.org/

Funded under the EU Sixth Framework Programme
Multimodal interfaces action line of the IST Programme
Integrated Projects
AMI (IST-506811) and AMIDA (IST–033812)

State of the Art Report

Recognizing Subjective Content in Text and Conversation

November 8, 2007

AMI Consortium State of the Art Report

Recognizing Subjective Content in Text and Conversation

November 8, 2007

Abstract

Applications such as meeting browsers and meeting assistants aim to identify, extract, and summarise *meeting content* — information about what happens and what is discussed in meetings. Most research in identifying and extracting meeting content has focused on primarily objective content, e.g., information about what topics are discussed and who is assigned to work on a given task. However, another type of meeting content that is important is the *subjective content* of meetings, i.e., the opinions and sentiments that the participants express during discussion in the meeting. Although there has been some work on recognizing subjective content in multiparty conversations, the majority of work in this area has focused on text. In this paper, we review the related work, both from text and from speech, that is relevant for the task of recognizing subjective content in meetings. We also present a new annotation scheme for marking subjective content in meetings.

1 Introduction

Applications such as meeting browsers and meeting assistants aim to identify, extract, and summarise *meeting content* — information about what happens and what is discussed in meetings. Some meeting content is primarily objective, for example, information about what topics are discussed [Hsueh and Moore, 2006] and who is assigned to work on a given task [Purver et al., 2006]. However, another type of meeting content that is important is the *subjective content* of meetings, that is, the opinions and sentiments that the participants express during discussion in the meeting. Recognizing subjective content is important because, intuitively, it seems that such information would help with existing meeting-browser tasks, such as decision detection [Hsueh and Moore, 2007]. But subjective content in and of itself is also interesting and important to extract and summarise. We would like to know not only what a particular decision was but who supported or opposed the decision. Imagine asking a meeting assistant not only to summarise the major ideas that were discussed but also the pros and cons expressed about those ideas.

To extract and summarise the subjective content of meetings, we first need to be able to identify when something subjective is being said and also to recognize the type of subjective content that is being expressed (e.g., positive or negative sentiment). However, to achieve the detailed analysis of subjective content that we would like, we also need to be able to identify the *source* and the *target* of the subjectivity—who the subjectivity is attributed to and what it is about. Although it is likely that most of the time the speaker is expressing his or her own opinions, it is not unusual for the speaker to report someone else's opinion or to be speaking on behalf of the group. For example, in (1) below, the speaker is reporting the opinion of the company, and in (2), the speaker is reporting information from a user study about remote controls. In example (3), the speaker is reiterating an opinion that the group as a whole holds.

> (1) The first one is that um uh the company's decided that teletext is outdated uh because of how popular the internet is.
>
> (2) Um people uh additionally aren't aren't liking the appearance of their products
>
> (3) Also we talked earlier about R_S_I and wanting to prevent um any sort of like Carpal Tunnely kind of thing

In the past few years, there has been some work on recognizing subjective content in multiparty conversations. For example, Wrede and Shriberg Wrede and Shriberg [2003a] have worked on recognizing meeting hotspots, which are a fairly coarse type of subjective content. Hillard et al. Hillard et al. [2003], Galley et al. Galley et al. [2004], and Hahn et al. Hahn et al. [2006] have worked on recognizing agreements and disagreements in meetings. Dialogue act coding schemes often include dialogue act tags for marking certain limited types of subjective content [Bhagat et al., August 2003]. Most recently, Somasundaran et al. Somasundaran et al. [2007b] worked to recognize utterances that express sentiment and arguing. While all of this research takes definite steps toward recognizing at least some aspect of the subjective content found in multiparty conversation, none of it provides both the level of detail and coverage of the subjective content that we believe is important to identify from meetings.

In contrast to the fairly limited amount of work on subjective content in meetings and conversation, the past few years have seen a surge of research in the recognition of subjective content in textual discourse. Annotation schemes have been proposed for marking opinions and other types of subjective content (e.g., Wiebe et al. [2005] and Martin and White [2005]), and corpora with detailed subjective content annotations have been produced. Researchers have worked on automatically identifying subjective sentences (e.g., Wiebe et al. [1999], Riloff and Wiebe [2003], and Yu and Hatzivassiloglou [2003]), recognizing the sentiment of phrases or sentences (e.g., Morinaga et al. [2002], Yu and Hatzivassiloglou [2003], Hu and Liu [2004], Popescu and Etzioni [2005], and Wilson et al. [2005]), recognizing expressions of opinions in context (e.g., Choi et al. [2006] and Breck et al. [2007]), and identifying who an opinion is attributed to (e.g., Bethard et al. [2004], Kim and Hovy [2004], and Choi et al. [2005]). There has also been a great deal of focus on automatically acquiring *a priori* subjective information about words and phrases, information which is then applied to automatically recognizing subjective content. This research includes learning words and phrases that are indicative of subjective language (e.g., Wiebe [2000], Riloff et al. [2003], Kim and Hovy [2005], Esuli and Sebastiani [2006]) as well as learning the polarity (semantic orientation) of words and phrases (e.g., Hatzivassiloglou and McKeown [1997], Turney and Littman [2003], Esuli and Sebastiani [2005], and Takamura et al. [2005]).

Monolingual text and multiparty conversation are very different types of discourse. For text, it is only the words on the page that convey whether or not something subjective is being expressed. In spoken conversation there are the words, as well as prosodic and visual cues that figure into the evidence to consider. However, given the depth of the research into recognizing subjectivity in text, exploring what approaches for text might also work for conversation is an obvious track to pursue.

With an eye toward our own goals of recognizing and extracting detailed subjective content in multiparty dialogue, in the first part of this paper we review some of the most relevant work on recognizing subjectivity in text. We start in Section 2 by giving an overview of the annotation schemes that have been developed for marking subjective content in text, and then in Section 3 we review the research in identifying subjective information about words and phrases. Finally, in Section 4 we review the research in automatic subjectivity and sentiment analysis in text that is most relevant to recognizing subjective content in conversation.

In the remaining sections, we focus on subjective content in speech and conversation. In Section 5 we give a brief overview of the research on emotion recognition, focusing on the work that has been done in spontaneous speech. Then in Section 6, we review the research that has been done so far on recognizing subjective content in multiparty conversation. Finally, in Section 7 we present our annotation scheme for marking subjective content in meetings.

2 Annotating Subjective Content in Text

There have been two detailed conceptualisations proposed for fine-grained analysis and annotation of subjective content in text, the Multi-perspective Question Answering (MPQA) Annotation Scheme [Wiebe et al., 2005] and Appraisal Theory [White, 2002, Martin and White, 2005]. The MPQA Annotation Scheme was developed for marking opinions and emotions in news articles. Appraisal Theory is a framework for analyzing evaluation and stance in discourse. Both representations are concerned with systematically identifying expressions that in context are indicative of subjective content.

This section gives an overview of both the MPQA Scheme and Appraisal Theory, as well as a brief review of the work in sentence-level subjectivity annotation.

2.1 MPQA Annotation Scheme

The MPQA Annotation Scheme is centred around the concept of *private state* [Quirk et al., 1985]. A private state is any internal mental or emotional state, including opinions, beliefs, sentiments, emotions, evaluations, uncertainties, and speculations, among others. In its most basic representation, a private state can be described based on its functional components: the state of an *experiencer* holding an *attitude* optionally toward a *target* [Wiebe, 1990, 1994].

The annotation scheme presented in [Wiebe et al., 2005] is a detailed, expression-level representation of private states and attributions that adapts and expands the more basic functional-component representation. The annotations in the scheme are represented as frames, with slots in the frames representing various attributes and properties. The initial MPQA scheme contains four annotation frames: **direct subjective frames**, **expressive subjective element frames**, **objective speech event frames**, and **agent frames**. In [Wilson, 2007], the MPQA scheme is extended to include two new types of annotation frames: **attitude frames** and **target frames**.

The direct subjective frame and the expressive subjective element frame are both used for representing private states, but they capture distinct ways that private states are expressed. Direct subjective frames are used to mark expressions that explicitly refer to private states and expressions that refer to speech events[1] in which a private state is expressed. The phrase "have doubts" in (4) is an example of an expression that explicitly refers to a private state. In (5), the phrase "was criticized" refers to a speech event in which a private state is being expressed, as does the phrase "said" in (6). The word "criticized" conveys that a negative evaluation was expressed by many people, even though their exact words are not given. With "said" in 6, it is the quoted speech that conveys the private state of the speaker, specifically the phrase "a breath of fresh air." Expressive subjective element frames are used to mark expressions that indirectly express private states, through the way something is described or through a particular wording. The phrase "a breath of fresh air" is an example of an expressive subjective element, as is the phrase "missed opportunity of historic proportions" in (7).

> (4) Democrats also have doubts about Miers' suitability for the high court.
> (5) Miers' nomination was criticized from people all over the political spectrum.
> (6) "She [Miers] will be a breath of fresh air for the Supreme Court," LaBoon said.
> (7) This the nomination of Miers is a missed opportunity of historic proportions.

Although private states are often expressed during speech events, not all speech events express private states. The objective speech event frame in the MPQA scheme is used to mark speech event phrases that refer to these objective speech events. In sentence (8), an objective speech event is marked on the word "said."

> (8) White House spokesman Jim Dyke said Miers' confirmation hearings are set to begin Nov. 7.

The agent frame in the scheme is used to mark noun phrases that refer to sources of private states and speech events. The source of a private state is the experiencer of the private state, and the source of a speech event is its speaker or writer. In (4) above, "Democrats" would be marked as an agent, as would "people all over the political spectrum" in (5) and "LaBoon" in (6).

All of the above annotation frames contain various attributes used to further characterize each expression that is annotated. Both private state frames, for example, include attributes for capturing the intensity of the private state being expressed and the polarity of the expression that is marked. One attribute that is included in all the annotation frames is

[1] A speech event is considered any event of speaking or writing.

Table 1: Attitude Types in the MPQA Scheme

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

the *nested source* attribute, which represents a key part of the MPQA annotation scheme. We describe this attribute below; details on the other frame attributes can be found in [Wiebe et al., 2005].

As previously mentioned, the source of a private state is the experiencer of the private state, and the source of a speech event is its speaker or writer. However, in textual discourse such as the news, there are frequently *layers of attribution*. For example, in (4) above, it is according to the writer of the sentence that the Democrats have doubts. Similarly, in (5) is it according to the writer that people are criticising the nomination. The *nested source* attribute captures these layers of attribution. In sentence (4), both the direct subjective frame ("have doubts") and the agent frame ("Democrats") are marked with the attribute $nestedsource = \langle writer, democrats \rangle$, where *writer* and *democrats* are unique identifiers that represent those agents in the discourse. Similarly, in (6) the expressive subjective element frame ("breath of fresh air"), the direct subjective frame ("said"), and the agent frame ("LaBoon") are all marked with the attribute $nestedsource = \langle writer, laboon \rangle$. In the example sentences above, there are no more than two layers of attribution; sentence (7) only has one layer for the writer of the sentence. However, in the news domain, it is not uncommon to find three or even more layers of attribution.

The last two types of annotation frames in the MPQA scheme are the attitude frame and the target frame [Wilson, 2007]. The attitude frames are linked to direct subjective frames. The purpose of an attitude frames is to capture the attitude being expressed overall by the private state to which it is linked. Similarly, target frames are linked to attitude frames; they are used to capture the target of the attitudes to which they are linked. The types of attitudes that are included in the attitude frame representation are listed in Table 1.

To date, the MPQA Annotation scheme has been used to annotate a corpus of 535 news articles (about 10,000 sentences) [2]. The MPQA annotations have been used in sentence-level subjectivity classification, phrase-level subjectivity and sentiment recognition, and source identification.

[2] Freely available at http://www.cs.pitt.edu/mpqa.

2.2 Appraisal Theory

Appraisal Theory [White, 2002, Martin and White, 2005] grew out of and seeks to extend the representation of language and meaning offered by Systemic Functional Linguistics (see Halliday [1985/1994]). The focus of Appraisal Theory is on analyzing how writers and speakers express attitude and stance, as well as how they position themselves with respect to their readers and listeners.

Figure 1, taken from [Martin and White, 2005] page 38, gives an overview of the taxonomy of Appraisal Theory. The Appraisal framework covers three main concepts, **Engagement**, **Attitude**, and **Graduation**. Engagement deals with what they call *intersubjective positioning*, which includes things like attribution and how the the writer positions himself or herself with respect to other viewpoints. Attitude is concerned with feelings and evaluations. This category further breaks down into **Affect**, **Judgment**, and **Appreciation**. Affect focuses on positive and negative feelings and emotions, Judgment is concerned with the evaluation of behavior, and Appreciation focuses on the evaluation of things. The last domain, **Graduation**, considers how attitudes are intensified or diminished, and how categories are sharpened (e.g., he's a *true* friend) and blurred (e.g., he's *sort of* a friend).

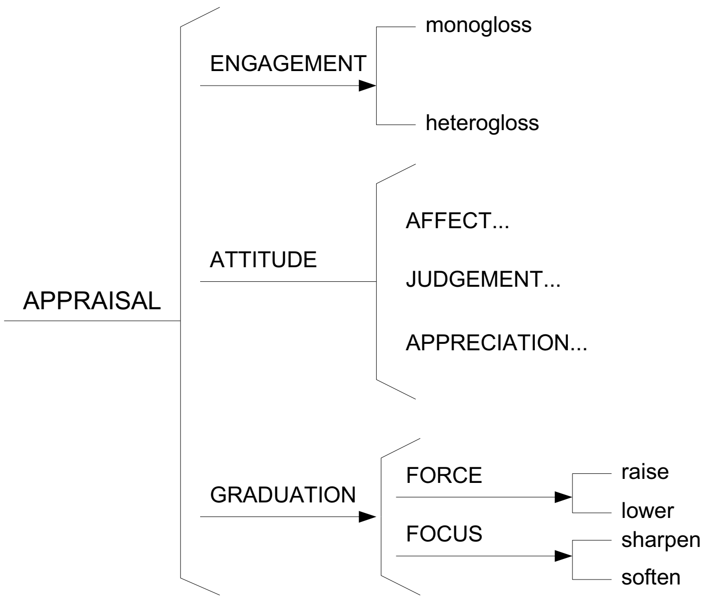
To date, Appraisal Theory has received only a limited amount of attention from the NLP community. Although it has been used to evaluate various types of discourse, including media commentary, casual conversation, and plays and literature, it has not yet been used to annotate large corpora, which could then be made available for exploration and evaluation using automatic methods. Recently, Read et al. [2007] began investigating whether the concepts and categories proposed by Appraisal Theory can be annotated reliably. In other work, researchers investigated whether lists of words, organized according to the Appraisal categories of Affect, Judgment, and Appreciation, were useful for the automatic classification of reviews [Whitelaw et al., 2005].

2.3 Other Subjective Content Annotations in Text

Aside from the MPQA Corpus and Appraisal Theory, annotation of subjective content in text has also been performed by Yu and Hatzivassiloglou [2003], Bethard et al. [2004], Kim and Hovy [2004], Hu and Liu [2004], Bruce and Wiebe [1999], and Wiebe et al. [2004]. The annotation schemes used by Bruce and Wiebe [1999] and Wiebe et al. [2004] are earlier, less detailed versions of the MPQA annotation scheme. Bruce and Wiebe perform sentence-level subjectivity annotations; the annotations in Wiebe et al. capture only expressive subjective elements.

The corpora developed by Yu and Hatzivassiloglou [2003], Bethard et al. [2004], and Kim and Hovy [2004] are annotated with sentence-level subjectivity and/or sentiment annotations. The corpus developed by Hu and Liu [2004] is a bit different from the others. They annotate targets, specifically products and product features in review data. However they do not 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. These annotations simply capture whether in a sentence there is a positive or negative sentiment toward a given target.

Figure 1: Overview of the Appraisal Theory taxonomy, from [Martin and White, 2005] page 38.



3 Learning Subjectivity Information about Words and Phrases

One aspect of subjectivity analysis that has received a fair amount attention is learning subjective words and phrases and learning the *polarity* or *semantic orientation* of words and phrases. This information is then typically compiled into a lexicon for use by systems seeking to recognise opinions and sentiments in context. Being able to automatically acquire information about the subjectivity and polarity of words is important for any system working with text or conversation that hopes to achieve good coverage in recognizing subjective content. People use an amazingly wide variety of language when expressing opinions and emotions. Systems that rely only on the words seen in annotated training data are unlikely to have enough knowledge to achieve the best results.

Researchers have explored various methods for learning the *a priori* subjectivity or polarity of words and phrases. Some exploit syntactic and semantic relationships that provide information about how two words are related in terms of their subjectivity or polarity. For example, we can infer subjective information about words that are joined by conjunctions. If one of the words in a conjunction is subjective, the other is likely to be subjective as well. Similarly, two words connected with the conjunction *and* are likely to have the same polarity, and words connected with the conjunction *but* typically have the opposite polarity. Semantic relationships like synonymy and antonymy provide similar sorts of information. If a word (or more specifically a word sense) is subjective, its synonyms and antonyms will be subjective too. Synonymy and antonymy also tell us whether certain words typically have the same or the opposite polarity.

Another common approach to learning subjectivity information about about words is by measuring how words pattern or associate statistically with known subjective/positive/negative words in a large corpus. These approaches work on the assumption that subjective words and words of the same polarity will be found near each other or will have similar distributions.

3.1 Exploiting Known Syntactic and Semantic Relationships

Hatzivassiloglou and McKeown [1997] were the first to use the co-occurrence of words in conjunctions to learn the polarity of words automatically. Their approach starts by extracting conjunctions of adjectives from a 21 million word news corpus and training a log-linear regression model to determine whether conjoined pairs of adjectives have the same or different polarity. Once they have this information, they use a clustering algorithm to separate the adjectives into positive and negative sets.

Kanayama and Nasukawa [2006] build on the ideas of Hatzivassiloglou and McKeown by considering what information about the polarity of words and phrases can be gleaned from discourse connectives and *context coherency*. Context coherency assumes that clauses with the same polarity will appear successively unless the context is changed with certain types of discourse markers. Kanayama and Nasukawa actually start with a fairly large, general-purpose collection of positive and negative words and phrases, with the goal of expanding their lexicon with positive and negative *domain dependent* words and phrases.

Kamps and Marx [2002] were the first to use semantic relationships to assign polarities to words automatically. Their approach uses synonymy links in the WordNet lexical database [Fellbaum, 1998] to determine the polarity of adjectives. Given an adjective *a* and two reference words that are antonyms (e.g., *good* and *bad*), Kamps and Marx compute whether *a* is more closely related through *SYNSET* (synonym set) links to the positive or to the negative reference word. Whichever reference word the adjective is closer to determines its polarity.

Since the work of Kamps and Marx, many other researchers have looked to WordNet and other thesauri for help learning the polarity of words. Hu and Liu [2004] and Kim and Hovy [2004] both start with small sets of positive and negative seed words and use synonymy and antonymy information from WordNet to grow these sets. Esuli and Sebastiani [2005] and Andreevskaia and Bergler [2006] also bootstrap from seed words, but their approaches make use of the glosses in WordNet as well as information about lexical relationships. Takamura et al. [2005] take a unique approach to learning the polarity words. They combine information from WordNet and information from corpora about the occurrence of words in conjunctions into *spin models*. A spin model models a set of electrons. In the models of Takamura et al., each electron corresponds to a word, and the up or down spin of the electron represents the word's polarity. The various types of information are represented as either same or different polarity links between electrons. To learn the polarity or spin of each word, Takamura et al. start by setting the polarity of a small set of seed words. This information is then propagated throughout the network until convergence is reached. Lexical relationships and glosses in WordNet have also been used to learn word subjectivity [Esuli and Sebastiani, 2006] and to assign words to categories from Appraisal Theory [Whitelaw et al., 2005]. There has also been work on automatically assigning subjectivity information to WordNet senses [Wiebe and Mihalcea, 2006, Esuli and Sebastiani, 2007].

3.2 Statistical Word Associations and Distributional Similarities

Several researchers have investigated using statistical measures of word association to predict the polarity or subjectivity of words. Turney and Littman [2003] use a modified version of Pointwise Mutual Information (PMI). For their corpus, they use the web. To predict the polarity of a given word, they start with small sets of positive and negative seed words and submit queries to the AltaVista search engine to see how many hits the target word has that are NEAR[3] the seed words. The polarity of the word is then determined by the seed set with which it has the highest PMI. Baroni and Vegnaduzzo [2004] take Turney and Littman's method and apply it to learning subjective words. Yu and Hatzivassiloglou [2003] measure positive and negative word associations using a modified log-likelihood ratio and a very large corpus of news articles.

Wiebe et al. [2004] hypothesized that subjective words could be expected to have similar patterns of distribution. To investigate this, they used Dekang Lin's Lin [1998] method for clustering words based on their distributional similarity to identify sets of subjective verbs and adjectives. The seed words for this process were the adjectives and verbs in editorials and other opinion-piece articles in the Wall Street Journal.

Riloff et al. [2003] and Riloff and Wiebe [2003] worked on learning subjective nouns and subjective extraction patterns. Extraction patterns are lexico-syntactic expressions that

[3] NEAR was an operator in the AltaVista search engine.

were originally developed for information extraction. As with others, Riloff et al. take a bootstrapping approach. Given a set of seed words that represent the semantic class of interest, in their case highly subjective nouns, their algorithms look for words that appear in the same extraction patterns as the seed words and determine which of these new words are the best to add to the set of seeds. The process then iterates. In [Riloff and Wiebe, 2003], Riloff and Wiebe switch their focus to identifying subjective extraction patterns.

Takamura et al. [2006] have also worked on identifying the polarity of phrases. Unlike the research above, their approach relies on hand-annotated data. Nevertheless, it is worth mentioning. Takamura et al. propose latent-variable models to capture the polarity of adjective-noun pairs. One variable corresponds to nouns and the other to adjectives. The data they use for both training and testing consists of a large collection of adjective-noun pairs, extracted from news data and hand annotated for their polarity. Interestingly, what they end up learning is often domain-dependent positive and negative phrases.

4 Automatic Recognition of Subjective Content in Text

Research on subjectivity analysis in text ranges from work on identifying the subjective information in words and phrases in context (e.g., Popescu and Etzioni [2005], Wilson et al. [2005], and Breck et al. [2007]), to work classifying the subjectivity of documents (e.g., Pang et al. [2002], Turney [2002], Dave et al. [2003] Pang and Lee [2005], and Ng et al. [2006]). Of this research, the work that is most similar to the type of analysis of multiparty conversation that we are aiming for is the research on sentence-level and phrase-level subjectivity analysis.

The simplest approaches to recognizing subjective content in text involve a straightforward lookup of terms from a subjectivity lexicon, taking into account the influence of negation. For example, Morinaga et al. [2002] and Yi et al. [2003] use detailed, hand-compiled lexicons of positive and negative words and phrases to identify opinions. Yu and Hatzivassiloglou [2003], Kim and Hovy [2004], and Hu and Liu [2004] classify the sentiment of sentences by averaging, multiplying, or counting the polarity of the words from the lexicon that appear in a sentence.

Many different machine learning learning approaches have been applied to recognising the subjectivity or polarity of sentences and phrases, from supervised learning using naive Bayes [Riloff et al., 2003, Yu and Hatzivassiloglou, 2003], support vector machines and boosting [Kudo and Matsumoto, 2004, Wilson et al., 2005, 2006, Somasundaran et al., 2007b, Furuse et al., 2007], conditional random fields [Mao and Lebanon, 2006, Breck et al., 2007], and structured linear classifiers [McDonald et al., 2006], to semi-supervised [Wiebe and Riloff, 2005, Gamon et al., 2005, Kaji and Kitsuregawa, 2006, Suzuki et al., 2006] and unsupervised techniques [Popescu and Etzioni, 2005]. Riloff et al. use a wide array of information, including counts of various types of subjective words and phrases, the presence of adjectives and certain other parts of speech, and the density of key subjective and objective words, to classify subjective sentences from the news. Yu and Hatzivassiloglou also classify subjective sentences from the news. They obtain their best results using n-grams and lists of positive and negative words. Kudo and Matsumoto investigate the use of dependency relations in classifying the polarity of sentences. Wilson et al.

explore the utility of a wide range of lexical, syntactic, and discourse features for phrase-level sentiment analysis. The task of Breck et al. is similar; they investigate phrase-level, subjective expression identification. Wilson et al. also experiment with classifying the intensity of sentences and clauses. Somasundaran et al. classify the attitude of sentences from the news and from a Web discussion board, and then investigate whether this information is useful for improving question answering. Furuse et al. develop a subjective sentence classifier to use as a component in an opinion search engine. Mao and Lebanon and McDonald et al. both investigate sentence-level sentiment classification as part of the larger task of classifying document sentiment. Gamon et al. and Mei et al. approach the problem of sentiment analysis as one of joint classification of topic and sentiment.

Several of the semi-supervised and unsupervised approaches are worth further mention. Wiebe and Riloff [2005] developed an approach that uses high-precision, rule-based, subjective and objective sentence classifiers to automatically build a large training corpus from unannotated data. Although the training set contains noise, the quality of the data is good enough that when used to train a supervised learner, the performance of the resulting classifier rivals that of a classifier trained on human-annotated data. Kaji and Kitsuregawa [2006] use a similar approach to automatically create a polarity-tagged corpus to use in training a classifier for sentence sentiment classification. They make use of high-precision linguistic patterns and certain HTML structures to build their training corpus automatically from the Web. Popescu and Etzioni [2005] use an unsupervised classification technique called *relaxation labeling* [Hummel and Zucker, 1983] to classify the polarity of select opinion phrases. They take an iterative approach, using relaxation labeling first to determine the polarity of the words, then again to label the polarities of the words with respect to their targets. A third stage of relaxation labeling then is used to assign final polarities to the words, taking into consideration the presence of other polarity terms and negation.

5 Emotion Recognition in Speech and Dialogue

An area of research that is very closely related to identifying subjective content and that has received a great deal of attention is the research on emotion recognition. Early research in emotion recognition focused on *acted* emotions. However, in recent years the focus has shifted to recognizing emotions in spontaneous speech and interactions. This later work is the research we overview in this section.

A number of different schemes have been proposed for representing and modelling emotion. Cowie and Cornelius [2003] give a good overview of the various models and taxonomies that have been proposed. Although some researchers propose fairly complex categorical schemes (e.g., Craggs and Wood [2004] and Devillers et al. [2005]), it is more common to find schemes that focus on just a few categories, for example, positive/negative/neutral (e.g., Litman and Forbes-Riley [2006], Neiberg et al. [2006], and Reidsma et al. [2006]) or negative/non-negative (e.g., Lee et al. [2002] and Shafran et al. [2003]). One reason for focusing on fewer rather than more emotion categories is the difficulty of the task. The more fine-grained the set of emotion categories, the harder the categories will be to recognize, both for human annotators and for automatic systems. In fact, even when an emotion annotation scheme has a larger set of fine-grained categories, researchers often end up conflating these into positive/negative or other more general categories for automatic classification experiments (e.g., Devillers et al. [2005]).

Researchers have applied any number of machine learning algorithms to the task of recognizing emotion, including decision trees, support vector machines, multi-layer perceptrons, Gaussian mixture models, boosting, and k-nearest neighbor. Although this research may suggest that certain approaches may be more useful than others for recognizing subjective content, the more valuable information to glean from the emotion recognition research is information about which features are the most promising. Prosodic and lexical features have of course been used for emotion classification, but other features have been found useful as well. For example, Devillers et al. [2005] and Forbes-Riley and Litman [2004] have found speech disfluencies to be useful. Forbes-Riley and Litman also found discourse information, such as the type of dialogue act in the previous turn to be informative.

6 Research in Recognizing Subjective Content in Multiparty Dialogue

6.1 Sentiment and Arguing Recognition

In recent work, Somasundaran et al. [2007a] developed an annotation scheme for marking expressions of sentiment and arguing in multiparty dialogue. They also conducted experiments in the automatic recognition of sentiment and arguing at both the sentence and turn levels.

The definitions for sentiment and arguing used by Somasundaran et al. in their annotation scheme were adapted from the attitude categories in [Wilson, 2007]. Sentiments include emotions, evaluations, judgments, feelings and stances. Arguing is defined as arguing for something or arguing that something is true. In the following examples (taken from [Somasundaran et al., 2007a]), the underlined words are considered arguing expressions.

(9) We ought to get this button

(10) Clearly, we cannot afford to use speech recognition

In their scheme, sentiment and arguing are not broken down into more fine-grained positive and negative categories.

Using their annotation scheme, Somasundaran et al. annotated 7 meetings from the AMI Meeting Corpus [Carletta et al., 2005]. Interannotator agreement ranges from 0.716 to 0.826 kappas at the turn level, and from 0.677 to 0.789 kappas at the sentence level.

To automatically recognize sentiment and arguing, Somasundaran et al. use support vector machines and perform experiments using 20-fold cross validation. The features they use include the words in the sentence or turn, counts of words from various word lists, and information about the flow of the discourse, represented using dialogue act and adjacency pair features. For sentiment recognition, positive and negative word lists from the General Inquirer [Stone et al., 1966] are used, as well as lists of strongly subjective words, weakly

	Baseline	Acc	Prec	Recall	F-measure
Arguing, turns	82.84	89.28	73.17	54.98	61.37
Arguing, sentences	85.50	90.30	73.22	51.32	59.20
Sentiment, turns	79.12	88.66	82.01	57.89	66.88
Sentiment, sentences	82.16	89.95	82.49	55.42	65.62

Table 2: Best results for sentiment and arguing classification reported by Somasundaran et al. [2007a]

subjective words, intensifiers, and valence shifters from [Wilson et al., 2005]. For arguing recognition, Somasundaran et al. compiled a list of arguing words and phrases through inspection (manual and semi-automatic) of both AMI meetings and meetings from the ICSI Meeting Corpus [Janin et al., 2003]. The dialogue acts within a sentence or turn are also used as features, as well as dialogue act–adjacency pair chains. For dialogue acts and adjacency pairs, they relied on manual annotations.

For both sentiment and arguing, their experiment using all the features produced the best results, although the majority of the gains come from the lexical features. Their results are summarised in Table 6.1. The baseline listed in the table for each experiment is the accuracy that results from choosing the most-frequent class. Although the precision is good, over 80% for sentiment, the difficulty of these tasks is revealed in the recall scores, the highest of which is only 58%.

6.2 Agreement and Disagreement

Hillard et al. [2003], Galley et al. [2004], and Hahn et al. [2006] have all worked on recognizing agreements and disagreements in multiparty conversation. Hillard et al. annotated the spurts[4] in 7 meetings from the ICSI Meeting Corpus [Janin et al., 2003] with one of four tags: *agreement*, *disagreement*, *backchannel*, and *other*. Frequent single-word spurts, such as *yeah* and *ok*, were not human annotated, but rather automatically separated out and categorized as backchannels. Hillard et al. report an inter-coder agreement 0.6 Kappa for tagging spurts with these categories. In the resulting annotations, agreements (9%) and disagreements (6%) are in the minority.

To recognise agreements and disagreements automatically, Hillard et al. train 3-way decision tree classifiers (the *agreement* and *backchannel* categories are merged) using both word-based and prosodic features. The word-based features include the total number of words in the spurt, the number of positive and negative keywords in the spurt, the class (agreement, disagreement, backchannel, discourse marker, other) of the first word of the spurt, which is determined using keywords, and the perplexity of the sequence of words in the spurt, which is computed using bigram language models for each of the four classes. Words with at least 5 instances and that have an *effectiveness ratio* > 0.6 are selected as keywords. Hillard et al. define the effective ratio as the frequency of a word in the desired class divided by the frequency of the word over all dissimilar classes combined. The bigram language models were trained in an unsupervised fashion by bootstrapping off of

[4] A spurt is a period of speech by one speaker that has no pauses of greater than one-half second.

the keywords. The prosodic features used by Hillard et al. include pause, fundamental frequency (F0), and duration, and features are generated for both the first word in the spurt and the spurt as a whole. In their experiments, the best classifier for hand-transcribed data uses only the keyword features and achieves an accuracy of 82% and a recall of 87% for combined agreements and disagreements (precision is not given). For ASR data, the best classifier uses all the word-based features and achieves an accuracy of 71% and a recall of 78%. Prosodic features do not perform as well as the word-based features, and when prosodic features are combined with the word-based features, there are no performance gains.

Galley et al. and Hahn et al. also use the data from the 7 ICSI meetings annotated by Hillard et al. with agreements and disagreements. Galley et al. investigate whether features capturing speaker interactions are useful for recognizing agreement/disagreement. For their approach, they model the problem as a sequence tagging problem using a Bayesian network and maximum entropy modelling to define the probability distribution of each node in the network. In addition to features capturing speaker interactions, they use lexical and durational features, which are similar to those used by Hillard et al. To identify speaker interactions, Galley et al. train a maximum entropy model to recognize adjacency pairs. In 3-way classification, Galley et al. achieve an accuracy of 86.92%, and for 4-way classification, they report an accuracy of 84.07%. As with Hillard et al., the lexical features prove to be the most helpful; adding durational features and features capturing speaker interactions gives only a slight boost to performance.

Hahn et al. investigate the use of contrast classifiers [Peng et al., 2003] for classifying agreements/disagreements. One challenge of classifying agreements and disagreements is the highly skewed distribution, with agreements and disagreements each making up only a small portion of the data. Contrast classifiers discriminate between labelled and unlabelled data for a given class. When a contrast classifier is trained for each class, only instances from a single class in the labelled data are used, and the data distribution within that class is modelled independently of the other classes. Because of this, a contrast classifier will not be as highly biased toward the majority class as classifiers trained over the imbalanced classes. The overall classifier that makes predictions in the test data is then an ensemble of contrast classifiers. In their experiments, Hahn et al. use only word-based features similar to those used by Hillard et al. Their best results are comparable to those achieved by Galley et al. However, the contrast-classifier approach gives only a slight improvement over straightforward supervised learning.

6.3 Hotspots in Meetings

Hotspots are places in a meeting in which the participants are highly involved in the discussion. Although high involvement does not necessarily mean there will also be subjective content, in practice, we expect more sentiments, opinions, and arguments to be expressed when participants are highly involved in the discussion.

Wrede and Shriberg [2003a,b] explore the recognition of hotspots in the ICSI Meeting Corpus. Rather than trying to define boundaries of hotspots, Wrede and Shriberg annotated individual utterances in terms of speaker involvement. Four categories were used: *amusement*, *disagreement*, *other*, and *not particularly involved.* Inter-annotator agreement for distinguishing the four categories was fairly low (0.48 kappa), with agreement for distinguishing just between involved and not involved being somewhat higher (0.59 kappa).

In [Wrede and Shriberg, 2003a], Wrede and Shriberg explore the correlation between involvement and a wide array of acoustic features. The features most strongly correlated with involvement were the maximums and averages of speaker-normalised fundamental frequency (F0). In [Wrede and Shriberg, 2003b], Wrede and Shriberg use hand-annotated dialogue acts to predict involvement.

6.4 Subjective Dialogue Acts

The dialogue act of an utterance refers to the intention of the speaker in speaking that particular utterance. Although dialogue act coding schemes vary, some schemes include labels specifically for marking when the intention of the speaker is to express something subjective. For example, the SWBD-DAMSL dialogue act coding scheme [Jurafsky et al., 1997] specifically includes a label for *Subjective Statements*. Other common labels for which we would expect the utterances marked to be subjective are *Suggestion* and *Assessment*).

The ICSI Meeting Corpus [Janin et al., 2003] and the AMI Meeting Corpus [Carletta et al., 2005] have both been annotated with dialogue acts, although the annotation schemes used are very different. The ICSI MRDA dialogue act coding scheme [Shriberg et al., 2004] uses a hierarchical organization of categories, with 11 general labels and 40 more specific, sub-category labels. The ICSI MRDA tagset includes *Assessment/Appreciation* and *Suggestion* labels. It also includes labels for which we would expect some, but not all, of the tagged utterances to be subjective: *Defending/Explanation*, labels in the *Responses* group (e.g., *Accept*, *Reject*, *Negative Answer*), and the labels in the *Politeness Mechanisms* group (e.g., *Sympathy, Apology*).

The AMI dialogue act coding scheme is made up of a much smaller set of labels than the ICSI MRDA scheme, only 15 labels in total. The AMI tagset also includes *Suggest* and *Assessment* labels. In addition, it includes the *Be Positive* and *Be Negative* labels. These tags are used to mark utterances in which the speaker's intention is to make an individual or the group feel more or less happy.

Although some subjective content is captured by specific dialogue act tags, other subjective content is not distinguished by the very nature of the dialogue act annotations. Dialogue acts mark the intention of the speaker. Thus, utterances in which the speaker reports about someone else's suggestions, assessments, and sentiments (e.g., sentences (1)–(3) above) will not be marked as such, because the speaker's intention for these utterances is to *inform*. Even for the speaker, while some types of subjective content correspond to typical dialogue act categories, other do not. Opinions, for example, may be *Assessments*, but they may be found in other types of dialogue acts as well.

6.5 Recognizing Emotionally Relevant Behaviour in Meetings

Laskowski and Burger [2006] propose an annotation scheme for marking what they call *emotionally relevant behavior* in the ISL Meeting Corpus [Burger et al., 2002]. Their

Discontent expressed in an attempt to slight
Other Discontent
Attempt to amuse
Acknowledgement or backchannel
Agreement expressed to improve another's self-esteem
Other Agreement
Confident Disagreement
Other Disagreement
Promotion of own ego
Doubt
Laughter
Proving or requesting information or opinion
Other

Table 3: Set of tags for marking emotionally relevant behavior in meetings

annotation scheme contains a total of 13 categories, which are listed in Table 6.5. To determine which category to apply to a speaker turn, annotators follow a decision tree with the categories making up the leaves in the tree.

In addition to the emotionally relevant behaviour categories, Laskowski and Burger also annotate turns with more general *positive*, *negative* and *neutral* emotion categories. For the more fine-grained scheme, agreement ranges from a 0.56 to 0.59 kappa. Agreement for the three-way emotion categories is 0.67 kappa.

Neiberg et al. [2006] use the ISL Corpus and the positive, negative, and neutral annotations in their emotion recognition experiments. For their experiments they use acoustic-prosodic features, specifically Mel-frequency Cepstral Coefficients (MFCCs) and pitch features, and lexical n-grams. Neiberg et al. report their highest accuracy for the experiment that uses all the features, however the highest recalls (0.57 average) are actually obtained using just the n-gram features.

6.6 Emotion Annotation of Meetings

Reidsma et al. [2006] and Jaimes et al. [2005] have also performed emotion annotation of meeting data. Reidsma et al. annotate the AMI Corpus by first having annotators segment the video of a person at the points where they perceive changes in the mental state of the person in the video. Once a meeting segment has been identified, the annotator characterises the segment in terms of its emotional polarity and intensity. The annotator may also choose to characterize the segment using one of fifteen mental-state labels, e.g., surprised, distracted, or amused.

Jaimes et al. Jaimes et al. [2005] experiment with labelling meeting videos in terms of polarity and intensity of emotion using continuous-scale labelling in real-time. They then investigate the relationship between the manual annotations and automatically extracted audio-visual features. Although their results are preliminary, they suggest correlations between posture changes and intensity of emotion in the meeting, and pitch and polarity of emotion.

7 AMIDA Scheme for Annotating Subjective Content in Meetings

Developing an annotation scheme for marking subjective content in meetings involves making several decisions. First, what type of subjective content would be most valuable to mark? To answer this question, it is important to consider the goals of the end application. Ideally, a meeting assistant would be able to extract and summarise information such as who supported or opposed a particular decision and what were the pros and cons behind a certain idea. To extract this kind of information the system will need to be able to identify positive and negative opinions, evaluations, and emotions, as well as agreements and disagreements. Although other types of subjectivity may also be informative, those listed above are the most important for our purposes. The meeting assistant will need to be able to differentiate between opinions belonging to the speaker and opinions being reported by the speaker that are attributed to someone else. Also important are the targets of opinions.

The next question to consider is what granularity of subjectivity annotation is most appropriate. Are expression-level annotations needed or would larger units such as turns be a better choice to annotate? The more fine-grained the annotations are, the better the subjective content is pinpointed. However, the more fine-grained and detailed the annotations are, the more time consuming they are to produce. Is it important or even feasible to mark the spans that refer to the sources and targets of opinions? Or, should source and target information just be captured as attributes on the subjectivity annotations? After exploring the meeting data and considering different levels of annotation, *utterance-level* annotations were decided on. For these annotations, *utterance* is defined loosely. An *utterance* may be a single phrase or expression, but whenever possible it is a sentence or proposition with the source and target of the subjectivity included in the span that is marked. Sources and targets are then marked as attributes of the subjectivity annotations.

In the first section below, we give an overview of the AMIDA annotation scheme. In developing the scheme, we adapted concepts from the MPQA Annotation Scheme [Wiebe et al., 2005, Wilson, 2007] to fit our research goals and to take into account the different nature of multiparty conversation. Recall that the MPQA Scheme was developed for annotating news articles. In Section 7.2, we report the results of an inter-annotator agreement study conducted to evaluate the reliability of the annotations.

7.1 Annotation Scheme

There are three main categories of annotations in the AMIDA scheme: *subjective utterances*, *objective polar utterances*, and *subjective questions*. Table 7.1 lists the annotation types in each category. The three main categories and the specific types of annotations in each category are described in more detail below.

Subjective Utterances
positive subjective
negative subjective
positive and negative subjective
uncertainty
other subjective
subjective fragment

Objective Polar Utterances
positive objective
negative objective

Subjective Questions
positive subjective question
negative subjective question
general subjective question

Table 4: AMIDA Subjectivity Annotation Types

7.1.1 Subjective Utterances

Formally defined, a *subjective utterance* is one in which a *private state* [Wiebe, 1990, 1994] is being expressed. At the minimum, a subjective utterance annotation spans the words and phrases being used to express the private state (either through word choice or prosody). However, if the source and/or target of the private state are referenced, they are also included in the span captured by the annotation.

The *positive subjective* annotation type is used to mark utterances expressing the following types of private states:

- positive sentiments (emotions, evaluations, and judgments)
- positive suggestions from which a positive sentiment can be inferred
- arguing for something
- beliefs from which a positive sentiment can be inferred
- agreements
- positive responses to subjective questions

Below are a few examples of various positive subjective annotations. The span of speech marked for each positive subjective annotation is in angle brackets.

> (11) And the other thing was that ⟨the company want the corporate colour and slogan to be implemented in the new design⟩.
> (12) So, like, ⟨I wonder if we might add something new to the to the remote control market, such as the lighting in your house⟩, or

> (13) Um ⟨so I believe the the advanced functions should maybe be hidden in a drawer, or something like tha from the bottom of it⟩.
> (14)
> A: Maybe like a touch screen or something
> B: ⟨Something like that, yeah⟩
> (15)
> B: Right, so do you think that should be like a main design aim of our remote control d you know, do your your satellite and your regular telly and your VCR and everything?
> D: ⟨I think so⟩. ⟨Yeah, yeah⟩.

The various negative private states included in the *negative subjective* annotation type are the opposite of the positive private states included in the positive subjective category:

- negative sentiments (emotions, evaluations, and judgments)
- negative suggestions from which a negative sentiment can be inferred
- arguing against something
- beliefs from which a negative sentiment can be inferred
- disagreements
- negative responses to subjective questions

Below are a few examples of negative subjective annotations.

> (16) ⟨Finding them is really a pain, you know⟩.
> (17) Um ⟨people uh additionally aren't aren't liking the appearance of their products⟩
> (18) Um I I haven't brought out one specific marketing idea, although my sense is that what we should try and think about is what are the current trends in materials and shapes and styles, and then use that. ⟨But not let that confine us technologically⟩.

The *positive and negative subjective* annotation type is for use in marking utterances where the positive and negative subjectivity cannot be clearly delineated. This happens with certain words and phrases that are inherently both positive and negative, for example, the word *bittersweet*. This can also happen when the grammatical structure makes it difficult to separate the positive and negative subjectivity into two utterances that clearly capture both the positive and the negative. There is an example of this in the the sentence below.

> (19) Um ⟨they've also suggested that we um we only use the remote control to control the television, not the VCR, DVD or anything else⟩.

The *uncertainty* and *other subjective* annotation types are included to capture utterances where other major types of private states are being expressed, even if those types are not the focus at this time. If these types of subjectivity are omitted, it would create a potential source of noise when it comes to recognizing automatically the types of subjectivity we are most interested in. This is also the reasoning for including the *subjective fragment* annotation type. Subjective fragments rarely have discernible content, but they are recognisably subjective and thus may be useful for learning subjective language.

7.1.2 Objective Polar Utterances

Objective polar utterances are statements or phrases that describe positive or negative factual information about something without conveying a private state. The sentence *The camera broke the first time I used it* gives an example of negative factual information; generally, something breaking the first time it is used is not good. An example of a sentence with positive factual information is *The camera lasted for several years past its warranty.*

Positive and negative factual information will often be part of an utterance that is subjective overall, either because of the way in which it is said (e.g., in an angry tone of voice) or because of the greater context. In such cases, the positive or negative factual information is not annotated. However, when positive or negative factual information is presented objectively, as in the following examples, it is marked as an objective polar utterance.

(20) Nobody uses teletext very much anymore (*negative objective*)
(21) Adults at least would pay more for voice recognition (*positive objective*)

Although objective polar utterances by definition are not subjective, they do contain positive and negative information that may be of interest to someone searching for sentiments and opinions in meeting data.

7.1.3 Subjective Questions

Subjective questions are questions where the speaker is eliciting the private state of someone else. In other words, the speaker is asking about what someone else thinks, feels, wants, likes, etc., and the speaker is expecting a response in which the other person expresses what he or she thinks, feels, wants, or likes. A subjective question may be a yes/no question, as in example (22) below, or it may be a more open-ended question, as in example (23).

(22) Do you like the large buttons?
(23) What do you think about the large buttons?

There are three types of subjective question annotations: *positive subjective question*, *negative subjective question*, and *general subjective questions*. Positive and negative subjective questions specifically are trying to elicit the positive or negative private state of someone else. For example, (22) above is a positive subjective question. General subjective questions are not slanted toward asking about a positive or negative private state. Question (23) above is an example of a general subjective question.

Subjective question are included in the annotation scheme for two reasons. First, because they use much of the same types of terminology that are used in subjective utterances (e.g., “like” and “think” in the examples above), they will be a source of noise when it comes to the automatic recognition of subjective content. Second, recognizing subjective questions may be important for identifying subjective utterances, because a subjective utterance is the expected response to a subjective question.

7.1.4 Sources

Each subjective utterance and objective polar utterance is marked with its *source*, who the private state or the objective information is attributed to. Below are the types of sources that can be marked on an annotation.

- Speaker
- Specific external entity (e.g., the company, speaker’s parents, UNICEF)
- General external entity (e.g., people, the man on the street)
- Other meeting participant
- Speaker speaking for group

7.1.5 Targets

Each subjective utterance and objective polar utterance is also marked with its *target*. In this annotation scheme, targets capture generally what the private state or the objective polar information is about.

- Remote design
- Remote design project
- Meeting Project
- Meeting
- Previous statement/idea
- Following statement/idea
- Speaker-self
- Other

The *remote design*, *remote design project*, and *meeting project* target types are task specific. In the meetings that are annotated, the participants play the part of a design team developing a new television remote control. Subjectivity expressed specifically about the design of the remote or remote controls in general is marked with the *remote design* target; subjectivity about other aspects of the project are marked with the *remote design project* target. At the end of the meetings in the scenario, the participants are asked to give a meta-evaluation of their meeting experience. These subjective expressions are marked with the *meeting project* target. The *Meeting* target type is used when subjectivity is expressed about the activity or the progress of the meeting itself. Subjectivity may also be marked as being about a *previous statement or idea* or about a *following statement or idea*. Finally, subjectivity may be self-directed (*speaker-self*).

7.2 Agreement Study

To evaluate whether the subjectivity annotations described above can be annotated reliably, two annotators independently annotated two meetings from the AMI corpus. Although annotations are marked on the meeting transcript, annotators were instructed to listen to the meeting audio and to view the meeting videos as part of the annotation process.

Because the annotators were choosing which spans to annotate rather than marking a fixed set of units, evaluating how well the two annotators agree is not straightforward. One possibility is to calculate precision and recall with respect to each annotator's tags. However, we found that only a small percentage of the subjectivity annotations marked by each annotator (13% for annotator A, and 27% for annotator B) actually cross dialogue act segment boundaries. Thus, we decided to measure agreement based on the dialogue act segments already marked in the corpus. This gives us the same set of units for each annotator, making for much easier calculation of agreement.

Because it is possible for a dialogue act segment to contain more than one subjectivity annotation, we measure agreement for each annotation type separately. Table 7.2 shows the agreement measured in terms of Kappa [Cohen, 1960] and percent agreement for the 1889 dialogue act segments marked in the two meetings used in the study. Agreement for whether a segment contains a subjective utterance is 0.56 kappa. The annotators have similar agreement for positive subjective utterances and subjective questions. Interestingly, agreement for whether a segment contains a negative subjective utterance is higher, 0.62 kappa, suggesting that negative subjectivity is easier to recognise, or at least less ambiguous, than positive subjectivity. Hypothesising that some of the disagreement might be due to confusion between the positive/negative subjective categories and the positive/negative objective categories, we also calculated agreement after conflating the two positive categories and the two negative categories. Although this did not lead to improved agreement for recognizing the combined positive categories, it did improve agreement for the combined negative categories, indicating that there is some confusion between negative subjective and negative objective utterances.

	Kappa	% Agreement
Subjective Utterances (excluding fragments)	0.56	79
Positive Subjective	0.58	84
Negative Subjective	0.62	92
Positive Subjective + Positive Objective	0.58	83
Negative Subjective + Negative Objective	0.68	93
Subjective Question	0.56	95

Table 5: Interannotator agreement for the AMIDA subjectivity annotations

References

Alia Andreevskaia and Sabine Bergler. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy, 2006.

Marco Baroni and Stefano Vegnaduzzo. 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, 2004.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In *Working Notes — Exploring Attitude and Affect in Text: Theories and Applications (AAAI Spring Symposium Series)*, 2004.

S. Bhagat, R. Dhillon, H. Carvey, and E. Shriberg. Labeling guide for dialog act tags in the meeting recorder meetings. Technical report 2, International Computer Science Institute, Berkeley, August 2003.

Eric Breck, Yejin Choi, and Claire Cardie. Identifying expressions of opinion in context. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, Hyderabad, India, 2007.

Rebecca Bruce and Janyce Wiebe. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2):187–205, 1999.

S. Burger, V. MacLaren, and H. Yu. The ISL Meeting Corpus: The impact of meeting type of speech style. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*, 2002.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI Meeting Corpus. In *Proceedings of the Measuring Behavior Symposium on "Annotating and Measuring Meeting Behavior"*, 2005.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 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, 2005.

Yejin Choi, Eric Breck, and Claire Cardie. 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, 2006.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

Roddy Cowie and Randolph R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1–2):5–32, 2003.

Richard Craggs and Mary McGee Wood. *Affective Dialogue Systems (Lecture Notes in CS Volume 3068/2004)*, chapter A Categorical Annotation Scheme for Emotion in the Linguistic content of Dialogue, pages 89–100. Springer Berlin/Heidelberg, 2004.

Kushal Dave, Steve Lawrence, and David M. Pennock. 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, 2003. Available at http://www2003.org.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18:407–422, 2005.

Andrea Esuli and Fabrizio Sebastiani. PageRanking WordNet synsets: An application to opinion mining. In *Proceedings of ACL-2007*, 2007.

Andrea Esuli and Fabrizio Sebastiani. 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, 2006. doi: http://acl.ldc.upenn.edu/E/E06/E06-1025.pdf.

Andrea Esuli and Fabrizio Sebastiani. 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, 2005.

Christiane Fellbaum, editor. *WordNet: An electronic lexical database*. MIT Press, Cambridge, 1998.

Kate Forbes-Riley and Diane J. Litman. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceedings of HLT/NAACL*, 2004.

Osamu Furuse, Nobuaki Hiroshima, Setsuo Yamada, and Ryoji Kataoka. Opinion sentence search engine on open-domain blog. In *Proceedings of IJCAI*, 2007.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 2004.

M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. In *Proceedings of the 6th International Symposium on Intelligent Data Analysis*, Madrid, Spain, 2005.

Sangyun Hahn, Richard Ladner, and Mari Ostendorf. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *Proceedings of HLT/-NAACL*, 2006.

M.A.K. Halliday. *An Introduction to Functional Grammar*. London: Edward Arnold, 1985/1994.

Vasileios Hatzivassiloglou and Kathy McKeown. 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, 1997.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT/NAACL*, 2003.

Pei-Yun Hsueh and Johanna Moore. Automatic topic segmentation and lablelling in multiparty dialogue. In *IEEE/ACM Workshop on Spoken Language Technology*, 2006.

Pei-Yun Hsueh and Johanna Moore. What decisions have you made: Automatic decision detection in conversational speech. In *Proceedings of HLT/NAACL*, 2007.

Minqing Hu and Bing Liu. 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, 2004.

R.A. Hummel and S.W. Zucker. On the foundations of relaxation labeling processes. *IEEE Transations on Pattern Analysis and Machine Intelligence (PAMI)*, 5(3):167–187, 1983.

Alejandro Jaimes, Takeshi Nagamine, Jianyi Liu, Kengo Omura, and Nicu Sebe. Affective meeting video analysis. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, 2005.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI Meeting Corpus. In *Proceedings of IEEE ICASSP 2003*, 2003.

D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL labeling project coder's manual, draft 13. Technical Report Technical Report 97-02, University of Colorado, Institute of Cognitive Science, 1997.

Nobuhiro Kaji and Masaru Kitsuregawa. 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, 2006.

Jaap Kamps and Maarten Marx. Words with attitude. In *1st International WordNet Conference*, pages 332–341, Mysore, India, 2002.

Hiroshi Kanayama and Tetsuya Nasukawa. 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, 2006.

Soo-Min Kim and Eduard Hovy. 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, 2005.

Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the Twentieth International Conference on Computational Linguistics (COLING 2004)*, pages 1267–1373, Geneva, Switzerland, 2004.

Taku Kudo and Yuji Matsumoto. 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, 2004.

K. Laskowski and S. Burger. Annotation and analysis of emotionally relevant behavior in the ISL Meeting Corpus. In *Proceedings of LREC*, 2006.

C. Lee, S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. In *Proceedings of ICSLP*, 2002.

Dekang Lin. 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, 1998.

Diane J. Litman and Kate Forbes-Riley. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590, 2006.

Yi Mao and Guy Lebanon. Isotonic conditional random fields and local sentiment flow. In *Proceedings of NIPS*, 2006.

J.R. Martin and P.R.R. White. *The Language of Evaluation: Appraisal in English.* Palgrave MacMillian, New York, N.Y., 2005.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of LREC*, 2006.

Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. 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, 2002.

Daniel Neiberg, Kjell Elenius, and Kornel Laskowski. Emotion recognition in spontaneous speech using GMMs. In *Proceedings of INTERSPEECH*, 2006.

Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. 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, 2006. URL `http://www.aclweb.org/anthology/P/P06/P06-2079`.

Bo Pang and Lillian Lee. 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, 2005.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 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, 2002.

K. Peng, S. Vucetic, B. Han, H. Xie, and Z Obradovic. Exploiting unlabeled data for improving accuracy of predictive data mining. In *Proceedings of ICDM*, 2003.

Ana-Maria Popescu and Oren Etzioni. 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, 2005.

M. Purver, P. Ehlen, and J. Niekrasz. Detecting action items in multi-party meetings: Annotation and initial experiments. In *Proceedings of MLMI*, 2006.

Randolph Quirk, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language.* Longman, New York, 1985.

Johathon Read, David Hope, and John Carroll. Annotating expressions of Appraisal in english. In *Proceedings of the First Linguistic Annotation Workshop (ACL-LAW)*, 2007.

Dennis Reidsma, Dirk Heylen, and Roeland Ordelman. Annotating emotion in meetings. In *Proceedings of LREC*, 2006.

Ellen Riloff and Janyce Wiebe. 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, 2003.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pages 25–32, Edmonton, Canada, 2003.

I. Shafran, M. Riley, and M. Mohri. Voice signatures. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H Carvey. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, 2004.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial 2007)*, 2007a.

Swapna Somasundaran, Theresa Wilson, Janyce Wiebe, and Veselin Stoyanov. 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*, 2007b.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA, 1966.

Yasuhiro Suzuki, Hiroya Takamura, and Manabu Okumura. 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, 2006.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 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, 2005.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. Latent variable models for semantic orientations of phrases. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, 2006.

Peter Turney. 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, 2002.

Peter Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

P.R.R. White. Appraisal: The language of attitudinal evaluation and intersubjective stance. In Verschueren, Ostman, blommaert, and Bulcaen, editors, *The Handbook of Pragmatics*, pages 1–27. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2002.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM-05)*, pages 625–631, 2005.

Janyce Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 735–740, Austin, Texas, 2000.

Janyce Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2): 233–287, 1994.

Janyce Wiebe. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text*. PhD thesis, State University of New York at Buffalo, 1990.

Janyce Wiebe and Rada Mihalcea. Word sense and subjectivity. In *Proceedings of COLING-ACL*, 2006.

Janyce Wiebe and Ellen Riloff. 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, 2005.

Janyce Wiebe, Rebecca Bruce, and Thomas O'Hara. 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, 1999.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39(2/3):164–210, 2005.

Theresa Wilson. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh, 2007.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 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)*, 2005.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99, 2006.

Britta Wrede and Elizabeth Shriberg. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proceedings of EUROSPEECH*, 2003a.

Britta Wrede and Elizabeth Shriberg. The relationship between dialogue acts and hot spots in meetings. In *Proceedings of the IEEE Speech Recognition and Understanding Workshop*, 2003b.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 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, 2003.

Hong Yu and Vasileios Hatzivassiloglou. 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, 2003.