AMIDA

Augmented Multi-party Interaction with Distance Access

`http://www.amidaproject.org/`

Integrated Project IST–033812

Funded under 6th FWP (Sixth Framework Programme)

Action Line: IST-2005-2.5.7 Multimodal interfaces

Deliverable D9.3: Compendium of State-of-the-Art reports

Due date: 01/10/2007 **Submission date:** 06/11/2007
Project start date: 1/10/2006 **Duration:** 36 months
Lead Contractor: USFD **Revision:** 1

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

D9.3: Compendium of State-of-the-Art reports

Abstract:

Using five separate reports, this document describes the "state-of-the-art" in AMIDA areas that were not covered within the previous AMI project. Four concern component technologies for meeting browsers and remote meeting assistants. They cover subjectivity, topic segmentation, summarization, and dialogue act recognition. One discusses meeting browsing technologies in general.

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).

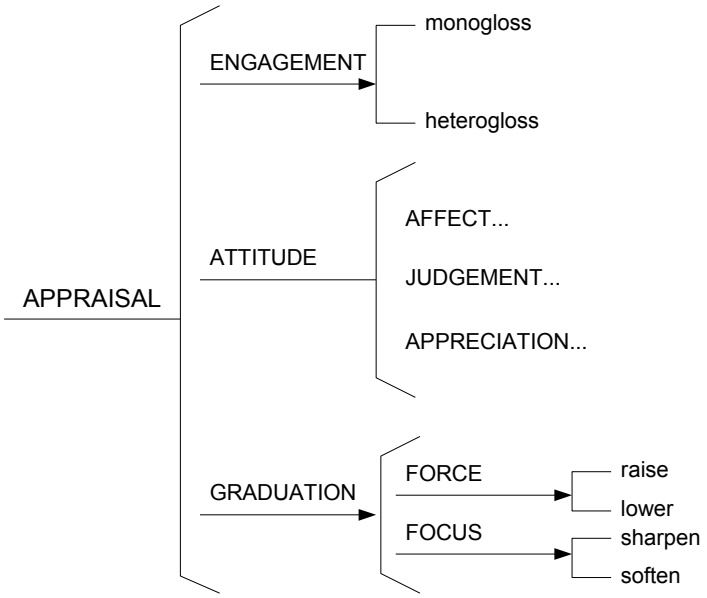
To date, Appraisal Theory has received only a limited amount of attention from the NLP community. Although it has been used to evaluate various types of discourse, including media commentary, casual conversation, and plays and literature, it has not yet been used to annotate large corpora, which could 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 agree-

ment 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.

Augmented **M**ulti-party **I**nteraction
http://www.amiproject.org

Augmented **M**ulti-party **I**nteraction with **D**istance **A**ccess
http://www.amidaproject.org

State-of-the-art overview

Recognition of Discourse Segments in Meetings

Updated version 10.10.2007

1 Introduction

Discourse is the deliberation process of what can be said about a specific topic. In common terms, it has been described in wikipedia as follows:

> Discourse is communication that goes back and forth (from the Latin, discursus, "running to and fro"), such as debate or argument. (c.f., wikipedia, as of 2007 [1])

Discourse prevails in our daily communication across widely ranging mediums such as written and spoken language. Regardless of the communication medium in use, our cognitive system can perform discourse segmentation effectively. This involves grouping coherent sequences of successive units (e.g., sentences, speech acts, or speaker turns) into discourse segments, each encompassing meanings beyond what is literally expressed in the individual units. Examples of our capacity to discourse segmentation are that we are able to interpret referring expressions, such as definite descriptions and pronouns, and resolve ellipsis. Also, we are good at summarising the gist of a particular segment and referring back to the relevant segment later.

In previous work of conversational discourse research, discourse segments are determined either from its informational content or by its intentional coherence. The former is done by grouping successive conversational units that are similar in the semantic focus of their expressions (that is, "what have been said") [33], or in the relations that link those units to each other, such as rhetorical predicates [27], coherence relations [37, 57], and conjunctive relations [33]. The latter is done by grouping units with similar expression styles (that is, "how the speaker expresses it")[2], or more generally, the underlying meaning and implicature of these successive units as a whole (that is, "what does the speaker imply when saying it") [25].

In fact, informationally and intentionally coherent segments are often posited as isomorphic, for example, in the hierarchical intention structure proposed in [30]. [53] have further provided empirical evidences on how most of the successive units that are intentionally similar are also informationally coherent.[3] Our cognitive system constantly monitors the phenomenon of informational coherence and that of intentional coherence in discourse – depending on the medium and the application, the system may choose to attend to one or both – to group successive units into discourse segments.

With respect to the different notions of discourse coherence, previous work has attempted different ways to determine discourse segments. On the one hand, informationally coherent conversational units have been captured by rendering inference-based formal methods (e.g. abduction) across propositional content in the discourse [55, 38]. These methods usually work as follows: Two adjacent units of discourse are considered at a time. If there exist a coherence relation (e.g., cause-effect, violated expectation, condition, similarity, contrast, elaboration, attribution, temporal sequence) [45, 71] between the situations described by the two units, then the two units can be concatenated as a coherent segment of discourse. Applying this algorithm successively to the whole discourse will result in a tree

[1] http://en.wikipedia.org/wiki/Discourse

[2] In [25]'s theory, we all have a "repertoire" we use to indicate different meanings. For example, I may always use the same gesture to let you know that I know what to do.

[3] [53] examined 183 sentences from general-interest magazines such as Reader's Digest. Despite that being informationally coherent (in this case, semantically cohesive) does not necessary translate to being intentionally coherent and, in converse, being intentionally coherent does not necessarily translate to being informationally coherent, most of the informationally coherent segments and intentionally coherent ones do correspond to each other.

structure for this discourse. However, the inference-based methods often assume the existence of full-fledged knowledge bases and, as a result, have problems with scalability. Therefore, past research has also explored other measures of informational coherence, for instance, semantic cohesiveness (i.e., a device that carries unity over text-like string representations of conversation) [53].

On the other hand, intentionally coherent segments have been determined through deriving discourse structure from its pragmatic context. Various theoretical models of discourse structure, for example, those based on individual-based speech acts [62, 26, 30] and on collaborative plans [63, 29], have been proposed. To understand the pragmatic characteristics of intentionally coherent successive units, empirical studies have been also conducted to analyze dialogue context in terms of both verbal features, e.g., discourse connectives [49, 4, 43], and non-verbal features, e.g., turn-taking cue [61, 47], acoustics (pitch range, contour, timing, energy level) [28], intonation pattern and speech rate [64], hand gesture, eye gaze, and head nod [13].

However, the theoretical and empirical studies on discourse coherence have focused more on the coverage of linguistic phenomenon rather than the computability of the proposed models and features. More recently, researchers have attempted to develop an automatic machinery to find discourse segment boundaries. A majority of works draw on the lexically cohesive characteristics of the segments to find informationally coherent segments. One successful approach is to view discourse segmentation as a time series problem amenable to signal processing. For example, TextTiling, an unsupervised lexical approach proposed by [35], looks for significant patterns in a quasi-temporal representation of the successive text units, and finds significantly disruptive patterns (i.e., where lexical cohesion scores change noticeably) that indicate a topic shift. [65] have extended the TextTiling approach to hypothesize segments in broadcast news.

Many other approaches view discourse segmentation as a dimension deduction problem similar to multinomial principal component analysis (PCA). On this front, variants of clustering algorithms have been proposed to group lexically similar units. In particular, Latent Semantic Analysis (LSA) [18], probabilistic Latent Semantic Analysis (pLSA) [39], and, more recently, Latent Dirichlet Allocation (LDA) [6] have been proposed to map lexical units to their associated semantic groups (a.k.a. topics). The representation of a discourse is then divided into major segments with respect to the semantic group features of these successive lexical units.

In practice, segmentation optimization can be achieved by using graph-cutting techniques to find segmentation that minimises inter-partition similarity without compromising intra-partition similarity [60, 58, 15, 69]. The graph-based techniques have been further attempted on hierarchical topic detection (HTD). It aims at organizing an unstructured news collection in a directed acyclic graph (DAG) structure, reflecting the topics discussed. (For more details, please refer to the report of the hierarchical topic detection task of TDT 2004 and [67].)

Segmentation optimization can also be achieved by applying the Hidden Markov model (HMM) and its variants (e.g., aspect HMM (AHMM)). The HMM-based framework consists of two major steps. First, k topic models (i.e., semantic groups) are constructed from large corpus (such as Wall Street Journal articles and CNN transcribed broadcasts), each model $T^{(j)}$, $1 \leq k$, referring to a smoothed language model of one semantically similar group found by some automatic clustering technique (e.g., K-means). Then, with respect to each of the identified k topic models, the probability of a given discourse unit being generated by this topic model is calculated. The topic of the highest probability will then be selected as the topic label (i.e. semantic group feature) of the unit. The observation of a discourse unit is considered as a collection of L mutually independent words that are generated by a topic model z_t. More formally, $o_t = w_{t,1}, w_{t,2}, w_{t,3}, ...w_{t,L}$. The emission probability can be computed

as follows:

$$P(o_t|z) = \prod_{n=1}^{L} P(w_i|z) \quad (1)$$

Transition probabilities among the topics and the self-loop probability are also calculated. Based on these probabilities, a search for the optimal segmentation will then be found by placing boundaries around where the associated topic of the current unit is different from that of the next unit [70, 5]. Various limitations of this supervised generative approach have been recognized. In particular, it requires sufficient labelled data for training representative topic models; the topics of a to-be-segmented discourse also have to fall within the range of the k topics which have associated models.

Finally, the machine learning approach has been further extended to combine cues that are central to the recognition of intentions and topical contents. Unlike previous works that use generative models, the intention-based segmentation works train discriminative models. Typically, in this framework the task is decomposed as a series of binary decisions: for each possible segment boundary site (i.e., the end of each discourse unit), the system extract the context of the site X. Given X, a pre-trained model $q(y|X)$ is then used to classify this site into a boundary class y, where $y \in YES, NO$. q can be learned from training data as a decision tree, i.e., a set of decision rules. For example, [28] and [49] have trained a decision tree to perform classification in spoken narratives, with respect to the acoustic contexts in discourse. q can also be a exponential model, i.e., a decision function which is parameterized by a set of weights for features in the context representation. [4] and [16] have achieved success on segmenting broadcast news by training exponential models with features that characterize both the information and the intentional coherence. The context X of each discourse unit is represented as a combination of these features, including the occurrence counts of topical words, that of discourse connectives in a neighbouring window, and the duration of pause.

2 Meeting Corpus

Spontaneous face-to-face dialogues in meetings violate many assumptions made by techniques previously developed for broadcast news (e.g., TDT and TRECVID), telephone conversations (e.g., Switchboard) [24] , and human-computer dialogues (e.g., DARPA Communicator) [20] . In order to develop techniques for understanding multiparty dialogues, smart meeting rooms have been built at several institutes to record large corpora of meetings in natural contexts, including ISL [8][4], CHIL ("Computers in the Human Interaction Loop"), LDC [17], NIST [22], ICSI [44], and in the context of the IM2/M4 project [51]. More recently, scenario-based meetings, in which participants are assigned to different roles and given specific tasks, have been recorded in the context of the CALO ("Cognitive Agent that Learns and Organizes") project (the Y2 Scenario Data) [9] and the AMI ("Augmented Multiparty Interaction") project [11].

the ICSI meeting corpus and the AMI meeting corpus, among the others, are the two corpora that contain discourse segmentation annotations. The ICSI meeting corpus (LDC2004S02) consists of the audio recording of seventy-five natural meetings in ICSI research groups. These meetings were recorded using close-talking far field head-mounted microphones and four desktop PZM microphones. The corpus includes manual orthographic transcriptions of all 75 meetings.

The AMI meeting corpus consists of the audio-video recordings of 173 meetings collected across

[4]The ISL Meeting Corpus contains 112 meetings collected at the Interactive Systems Laboratories at CMU during the years 2000-2001. The recorded meetings were either natural meetings, or artificial meetings, which were designed explicitly for the purposes of data collection but still had real topics and tasks. The duration of the meetings in this corpus ranges from eight to 64 minutes and averages at 34 minutes.

three sites, IDIAP, U of Edinburgh and TNO. This corpus also includes high quality, manually produced orthographic transcription for each individual speaker. It is different from the ICSI meeting corpus in several aspects. First, while all of the ICSI meetings are natural group meetings where participants needed to meet in real world, only 33 meetings of the AMI meetings are natural ones. Approximately two-thirds of AMI meetings (140 out of 173) are driven by a scenario, wherein four participants play the role of the project manager, marketing expert, industrial designer, and user interface designer in a design team, taking a design project from kick-off to completion. Second, in addition to audio recordings, the AMI meetings also come with video recordings recorded by individual and room-view video cameras, slides from a slide projector, the note-taking pen inputs, and input from an electronic whiteboard.

2.1 Structural Discourse Segmentation Annotation

One third of the ICSI meeting corpus (25 out of 75) comes with annotations of discourse segmentation.[5] The AMI project team have also produced discourse segmentation annotations for both the whole ICSI and AMI corpus. In these annotations, topic segmentation is used as a covering term of discourse segmentation, without differentiating information and intentional coherence. Annotators have the freedom to mark a topic as subordinated[6] wherever appropriate. Three human annotators used a tailored tool to perform topic segmentation in which they could choose to decompose a topic into subtopics, with at most three levels in the resulting hierarchy.

As it is expected that the preferred segmentation algorithm for predicting segment boundaries at different levels of granularity would be different, this research flattens the subtopic structure and consider only two levels of segmentation–top-level topics (TOP) and all subtopics (ALL). The top level of the structure signals either major topic shifts in discourse structure or serious abruption of the ongoing discussions. The second level of the structure signifies either a temporary digression or a discussion that is more focused on one aspect of the current major topic. Basic statistics of the topic segmentation annotations are reported in Table1. Compared to the ICSI corpus, the segmentation structure of the AMI corpus is much more shallower, with smaller difference between the number of TOP segments and that of ALL segments.

Take the topic segmentation annotation of a 60 minute meeting Bed003 in the ICSI corpus for example. In this meeting, the research team are discussing about the planning of an automatic speech recognition project. Four major topics, from “opening” to “general discourse features for higher layers” to “how to proceed” to “closing”. Depending on the complexity, each topic can be further divided into a number of subtopics. For instance, “how to proceed” can be subdivided to 4 subtopic segments, “segmenting off regions of features”, “ad-hoc probabilities”, “data collection” and “experimental setup”.

Average	TOP	ALL	Length
ICSI	6.96	17.2	40 mins
AMI	7.67	13.65	28 mins

Table 1: *Basic statistics of discourse segmentation annotations in the ICSI and the AMI corpus.*

Previous works have examined the reliability of human discourse segmentation annotations. [50] have reported that human annotators mostly agree with each other in the text segment boundaries

[5] In this annotation, Michel Galley et al.[21] have gathered together the majority codings from at least three coders per observation.

[6] In the AMI annotation, the subordinated topics can go down to two levels, while in the ICSI annotation, they can only go down to one.

they chose despite a margin of a few utterances. [54] have demonstrated the level of reliability of human segmentation annotations in spoken narratives is within a reasonable range.[7]

To establish reliability of the annotation procedures used for segmenting the meeting corpora, kappa statistics [10] have been calculated as a measurement of the agreement between the annotations of each pair of coders. We also reported on the overall segmentation error rate, Pk and WD. Pk [4] is the probability that two utterances drawn randomly from a document (in our case, a meeting transcript) are incorrectly identified as belonging to the same topic segment. WindowDiff (Wd) [56] calculates the error rate by moving a sliding window across the meeting transcript counting the number of times the hypothesized and reference segment boundaries are different.

Table 2 shows the average kappa statistics of the three pairs of coders on the top-level and sub-level segmentation respectively. [31] have reported kappa (pk/wd) of 0.41 (0.28/0.34) for determining the top-level and 0.45(0.27/0.35) for the sub-level segments in the ICSI meeting corpus. [42] have reported that the human annotators have achieved $\kappa = 0.79$ agreement on the TOP segment boundaries and $\kappa = 0.73$ agreement on the ALL segment boundaries. Do the kappa values shown here indicate reliable intercoder agreement? In computational linguistics, kappa values over 0.67 point to reliable intercoder agreement. But [19] have found that such interpretation does not hold true for all tasks. However, the low disagreement rate among codings in terms of the PK and WD scores can be used to argue for the reliability of the annotation procedure used in these studies.

Intercoder	Kappa	PK	WD
ICSI(TOP)	0.41	0.28	0.23
ICSI(SUB)	0.45	0.27	0.35
AMI (TOP)	0.66	0.11	0.17
AMI (SUB)	0.59	0.23	0.28

Table 2: *Intercoder agreement of annotations at the top-Level (TOP) and sub-Level (SUB) segments.*

A complete manual topic segmentation has been annotated for the ICSI meeting corpus and the AMI meeting corpus. In the ICSI corpus, topic labels were essentially free format. Annotators were asked to provide a free text label for each topic segment; they were encouraged to use keywords drawn from the transcription in these labels. However, to impose some level of consistency, some standard labels are also provided for annotating the off-topic discussions, such as “opening” and “chitchat”.

As for those AMI meetings that are scenario-driven, annotators are expected to find that most of the topics do recur. Therefore, they are given a standard set of topic descriptions that can be used as labels for each identified topic segment. Annotators will only add a new label if they cannot find a match in the standard set. The standard set of topic descriptions has been divided to three categories:

- Top segments refer to topics whose content largely reflects the meeting structure (e.g, presentation, discussion, evaluation, drawing exercise) and the key issues of the design task (e.g., project specs, user target group).
- ALL segments refer to parts of the top-level topics (e.g., project budget, look and usability, trend watching, components, materials and energy sources).
- Functional segments are those parts of the meeting that refer to either the varying process and flow of the meeting (e.g., opening, closing, agenda/equipment issues), or are simply irrelevant (e.g., chitchat).

[7] Seven annotators worked on segmenting the corpus, which consists of 20 narratives monologues about the same movie, taken from [14].

In addition to the manual transcriptions, these meeting corpora also come with ASR transcriptions. The ASR transcriptions were produced by [32], with an average WER of roughly 30%. The system used a vocabulary of 50,000 words, together with a trigram language model trained on a combination of in-domain meeting data, related texts found by web search, conversational telephone speech (CTS) transcripts and broadcast news transcripts (about 10^9 words in total), resulting in a test-set perplexity of about 80. The acoustic models comprised a set of context-dependent hidden Markov models, using gaussian mixture model output distributions. These were initially trained on CTS acoustic training data, and were adapted to the ICSI meetings domain using maximum a posteriori (MAP) adaptation. Further adaptation to individual speakers was achieved using vocal tract length normalisation and maximum likelihood linear regression. A four-fold cross-validation technique was employed: four recognizers were trained, with each employing 75% of the meetings as acoustic and language model training data, and then used to recognise the remaining 25% of the meetings.

3 Evaluation Metrics

3.1 Automatic Discourse Segmentation

To evaluate the performance of segmentation models, various metrics have been proposed in the field of text segmentation. The most typical example is accuracy. Previous work has shown that when class distributions display a high level of entropy, i.e. $P(c_i \mid T) \approx P(c_j \mid T), i \neq j$ for any two classes c and training data T, accuracy is an acceptable measure of quality for a classifier. But when class distributions are highly skewed, recall, precision and harmonic means of these like the F_β-score are better measures.

In fact, discourse segmentation is a typically class-imbalanced task. The number of linguistic units on which segmentation is based (like sentences) typically by far exceeds the number of actual topics. Consequently, optimizing a classifier for accuracy would automatically favor a majority classifier that labels all sentences as not initiating a new segment. Optimization for the classical notions of recall and precision would not work well here either: for instance, a discourse segmenter that always predicts a segment boundary close but not exactly corresponding to the ground truth prediction would produce zero recall and precision, while its performance can actually be quite good.

In respond to this problem, P_k and W_d were designed to overcome the limitations inherent in the use of precision and recall for discourse segmentation. [4] has defined the P_k measure as the probability that a randomly drawn pair of utterances are incorrectly predicted as coming from the same segment. Also, [56] have analyzed several weaknesses of the P_k measure and proposed an adapted metric WindowDiff (W_d). W_d is computed as the probability that the number of hypothesized and reference segment boundaries in a given window frame are different.

However, these specific measures like Pk and WindowDiff ([**?**]) compute recall and precision in a fixed-size window to alleviate this problem, but they do not penalize false negatives and false positives in the same way. For topic segmentation, false negatives probably should be treated on a par with false positives, to avoid undersegmentation. Recently, [23] proposed a new, cost-based metric called Pr_{error}:

$$Pr_{error} = C_{miss} \cdot Pr_{miss} + C_{fa} \cdot Pr_{fa} \tag{2}$$

Here, C_{miss} and C_{fa} are cost terms for false negatives and false alarms; Pr_{miss} is the probability that a predicted segmentation contains less boundaries than the ground truth segmentation in a certain interval of linguistic units (like words); Pr_{fa} denotes the probability that the predicted segmentation

in a given interval contains less boundaries than the ground truth segmentation. We refer the reader to [**?**] for further details and the exact computation of these probabilities.

3.2 Automatic Discourse Labelling

To evaluate the automatically generated labels against reference labels in the meeting corpus, relevant candidate metrics can be found in the fields of "story boundary detection" studied in TDT [66], TRECVID [46], and summarization studied in DUC [34]. Since the discourse segmentation annotators of some of the meeting corpus (e.g., the ICSI corpus) are free in their choice of keywords for topic labels, automatic evaluation of topic label assignment is difficult nad has not been attempted. For those meeting corpus (e.g., the AMI meeting corpus) that have their discourse labels selected from a predetermind set, overall classification accuracy is calculated as f-score (F1) to evaluate the performance of discourse labelling components [41]. We loop over each discourse segment in the standardized set. For each label in a predertermined set, precision is then computed as the total number of the discourse segments that have been assigned correct labels divided by the total number of discourse segments in the groud truth data; Recall is computed as the total number of the discourse segments that have been assigned correct labels divided by the number of segments that have been hypothesized as this label. A psuedo algorithm is given as below.

1. Loop(1) over each topic in the predetermined set
2. recall= total number of segments that have been assigned correctly to this topic/ total number of reference segments of this topic
3. precision= total number of segments that have been assigned correctly to this topic/total number of segments hypothesized as this topic
4. End Loop(2)

4 Recognition of Discourse Segments in Meetings

The problem of how to divide unstructured meeting speech into a number of locally coherent segments is important for two reasons: First, empirical analysis has shown that annotating transcripts with semantic information (e.g., topics) enables users to browse and find information from multimedia archives more efficiently [2]. Second, because the automatically generated segments make up for the lack of explicit orthographic cues (e.g., story and paragraph breaks) in conversational speech, dialogue segmentation is useful in many spoken language understanding tasks, including anaphora resolution [30], information retrieval (e.g., as input for the TREC Spoken Document Retrieval (SDR) task), and summarization [72].

As mentioned in Section 1, previous works have adopted three major approaches to tackle the problem of discourse segmentation: lexical-cohesion based approaches, topic modelling approaches, and supervised learning approaches. The first two can be operated in an unsupervised fashion. In the field of meeting discourse segmentation, [21] have extended the lexical cohesion-based TextTiling approach (named as LCSeg), and [59] have adapted the topic modelling approach to combine different topics so as to make this approach generalize well to segment meetings.

[21] has also applied the supervised learning approach to combine the outputs from LCSeg, which indicate information coherence, and other conversational features, which indicate speaker intentions.

Results have shown that the latter approach which trains a segmentation model with features that are extracted from knowledge sources beyond words, such as speaker interaction (e.g., overlap rate, pause, and speaker change) can outperform LCSeg. In addition, [3] have also pointed out that, when participant behaviours, e.g., note taking cues, are aggregated into the segmentation model, the performance can be further improved.

[41] have extended the supervised learning work in two ways: First, to understand whether there exists a difference in the preferred approach for predicting topic segmentation at different levels of granularity, it applied approaches that have been proposed for predicting granular-level topic shifts to the problem of identifying segments at a finer level. Second, as perfect human transcripts are always available, it has explored the impact on performance of using ASR output as opposed to human transcription.

The examination of the effect of features on performance shows that predicting top-level and predicting subtopic boundaries are two distinct tasks: (1) the lexical cohesion-based approach alone can capture the finer-level topic shifts, (2) the supervised learning approach, which combines lexical cohesion and intention-indicative features, performs better on predicting granular-level segments than on finer-level ones, and (3) applying feature selection, such as filtering cue phrase features with statistical metrics, can improve the performance of (2) on predicting finer-level segments by 10.46%. The examination of the effect of ASR transcripts has shown that despite the inevitable errors in ASR transcriptions, the preferred approach for predicting granular-level and finer-level segments does not change.

The experiments of [21] and [41] are both run on the ICSI corpus (LDC2004S02) [44]. [40] have applied these approaches on the AMI corpus [12]. However, results have shown that LCSeg is less successful in identifying "agenda-based conversation segments" (e.g., presentation, group discussion) in the AMI meetings. This is not surprising since LCSeg considers only lexical cohesion, and agenda-based segments are typically signalled more by intentional coherence than by informational coherence.

In many other researches which consider segmentation, a variety of features have been identified as indicative of segment boundaries in different types of recorded speech. For example, [7] have shown that a discourse segment often starts with relatively high pitched sounds and ends with sounds of pitch within a more compressed range. [54] have identified that topic shifts often occur after a pause of relatively long duration. Other prosodic cues (e.g., pitch contour, energy) have been studied for their correlation with story segments in read speech [68, 48, 16] and with theory-based discourse segments in spontaneous speech (e.g., direction-given monologue) [36]. In addition head and hand/forearm movements are used to detect group-action based segments [52, 1].

Therefore, [40] have further extended previous work to combine more features that can be extracted from dialogue contexts and multimedia inputs. Results have improved on previous work by 8.8% for granular-level segmentation and 5.4% for finer-level segmentation. Analysis of the effectiveness of the various features shows that lexical features (i.e., cue words) are the most essential feature class to be combined into the segmentation model. However, lexical features must be combined with other features, in particular, conversational features (i.e., overlap, pause, speaker activity change), to train well performing models. Furthermore, the multimodal features are essential to achieve good performance of a combined model. This is mainly because (1) the presence of the non-verbal features in the model can balance the tendency of models trained with lexical cues alone to over-predict, and (2) there is an interaction effect between these non-verbal features.

5 Application

The application needs of meeting speech segmentation is two-fold: On the one hand, the recognized discourse segments in meeting discourse form a quasi-summary for what have been transpired in a meeting and, in turn, provide the right level of details for users to interpret what the interlocutors are talking about in a meeting. Imagine the scenario that an industrial designer has missed a meeting and wanted to review the design team's discussion about the target user group. If the system can provide a discourse segment structure as shown in Figure 1, the users can then efficiently locate relevant information they are looking for (in this case, the segment about "target user group") from the list of segments. As evidenced in [2], discourse information does enable users to browse and find information from a meeting archive more efficiently. Moreover, when a recorded meeting has to be displayed on a mobile device, the recognised discourse segments can be used to construct an easy-to-grasp, thumb-nail view of the meeting. In short, discourse segmentation recognition has great potentials to enhance the current user interaction scheme of browsing and search.

On the other hand, discourse segment recognition benefits the development of other downstream meeting understanding applications. These applications include anaphora resolution [30], information retrieval (e.g., as input for the TREC Spoken Document Retrieval (SDR) task), summarization [72], and question answering. Take the application that needs to recover information for user queries for example, the recognized discourse segments can be used to guide the search of answer candidates toward those segments that are of topical relevance to the queries; the topical focuses of these segments can also serve as a means to rank the relevance of a list of answer candidates. The benefits of discourse segmentation on these applications would be even more evident when these applications have to be operated in an unfamiliar domain or in a foreign language environment.

6 Conclusion

We provided an overview of research in the area related to the recognition of discourse segmentation in meetings. The analyses concerns (1) the different notions of coherence central to discourse segmentation, (2) the features characteristics of coherence or the abruption of coherence, (3) the methods effective for finding discourse segments in text and spoken narratives, and (4) whether these methods and features can be effective for finding discourse segments in meetings.

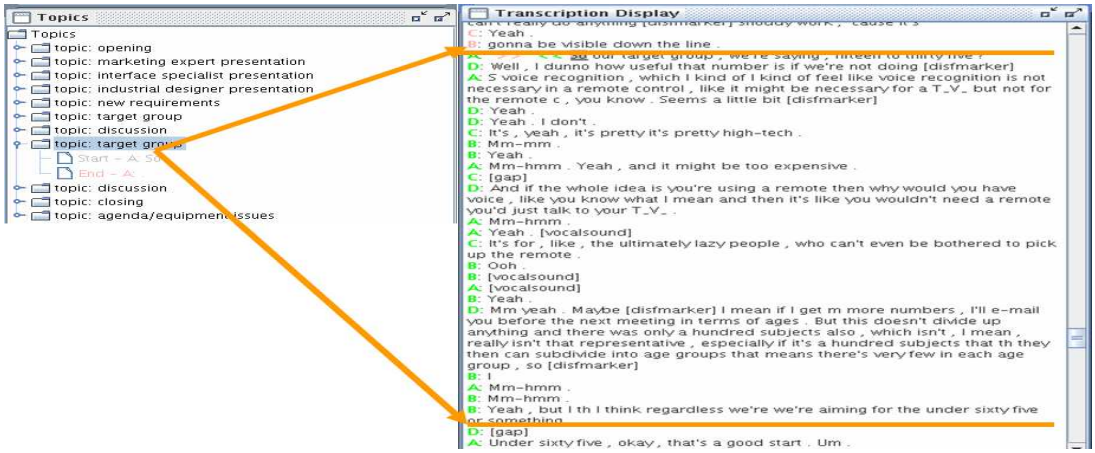


Figure 1: *Example of topic segmentation in a produce design meeting.*

The recognition of discourse segments requires recognition and tracking of lexical cohesion and other intention-indicative contexts, such as gesture/head movements, pitch, energy, rate of speech, and pause. There are many more works about inferring speaker intentions, for example, those in the line of human computer interaction research, we did not mention in this report. This is because our focus is on discourse segmentation. So we have only reviewed the discourse researches that are relevant to the recognition of segment boundaries.

Although recent research that used supervised learning approaches to combine various lexical cohesion and intention-indicative features have achieved success, it has at least two shortcomings: First, although these features are expected to be complementary to one another, few of them have studied how to systematically model the correlation among features in machine learned models. This has pointed out some possible improvements on applying some more sophisticated machine learning approaches, such as Conditional Random Fields, to overcome this shortcoming.

Second, training a well-performing discriminative model requires plentiful labelled data; yet, it is uncertain whether the trained model can be applied to segment meetings in a domain different from the labelled data. One solution is to apply unsupervised approaches. However, previous works in unsupervised meeting segmentation focus mainly on modelling word-related phenomenon, such as lexical cohesion and topical focus. Yet, we have seen in the supervised learning work that many other features beyond words, such as multimodal and dialogue contexts, are central to meeting segmentation. This is partially because meeting dialogues are spontaneous conversations in a multiparty environment, and naturally we have more communicative channels, such as body language, gaze engagement, gesture, and prosody, we can use to signal what we mean.

This has indicated the need of further investigation into how to combine multiple knowledge sources into the unsupervised approaches for meeting segmentation. To adaptively generalize the word-based approaches to combine multimodal features, two possible directions have thus arsed: (1) a more thorough empirical study about the synchronism mechanism between the intention-indicative features and the words, and (2) some novel ways to combine features in the current unsupervised segmentation approaches are also necessary.

References

[1] M. Al-Hames, A. Dielmann, D. GaticaPerez, S. Reiter, S. Renals, and D. Zhang. Multimodal integration for meeting group action segmentation and recognition. In *Proc. of MLMI 2005*, 2005.

[2] S. Banerjee, C. Rose, and A. I. Rudnicky. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proc. of the International Conference on Human-Computer Interaction*, 2005.

[3] S. Banerjee and A. Rudnicky. Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *Proc. of IUI 2006*, 2006.

[4] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34:177–210, 1999.

[5] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2001.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] G. Brown, K. L. Currie, and J. Kenworthe. *Questions of Intonation*. University Park Press, 1980.

[8] S. Burger, V. MacLaren, and H. Yu. The isl meeting corpus: The impact of meeting type on speech style. In *Proceedings of the ICSLP 2002*, 2002.

[9] CALO. Cognitive agent that learns and organizes. *http://www.ai.sri.com/project/CALO*, 2006.

[10] J. Carletta. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[11] J. Carletta et al. The AMI meeting corpus: A pre-announcement. In *Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2005.

[12] J. Carletta et al. The AMI meeting corpus: A pre-announcement. In S. Renals and S. Bengio, editors, *Springer-Verlag Lecture Notes in Computer Science*, volume 3869. Springer-Verlag, 2006.

[13] J. Cassell, Y. Nakano, T. Bickmore, C. Sidner, and C. Rich. Non-verbal cues for discourse structure. In *Association for Computational Linguistics Annual Conference*, 2001.

[14] W. L. Chafe. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Ablex Publishing Corporation, 1980.

[15] F. Choi, P. Wiemer-Hastings, and J. D. Moore. Latent semantic analysis for text segmentation. In L. Lee and D. Harman, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 109–117, 2001.

[16] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. Maximum entropy segmentation of broadcast news. In *Proc. of ICASP*, Philadelphia USA, 2005.

[17] C. Cieri, D. Miller, and K. Walker. Research methodologies, observations and outcomes in conversational speech data collection. In *Proceedings of the Human Language Technologies Conference (HLT)*, 2002.

[18] S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41 (6):391–40, 1990.

[19] B. Di Eugenio and M. G. Glass. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101, 2004.

[20] M. Eskenazi, A. Rudnicky, K. Gregory, P. Constantinides, R. Brennan, C. Bennett, and J. Allen. Data collection and processing in the carnegie mellon communicator. In *Proceedings of Eurospeech*, 1999.

[21] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proc. of ACL 2003*, 2003.

[22] J. S. Garofolo, C. D. Laprun, M. Michel, V. Stanford, and E. Tabassi. The NIST meeting room pilot corpus. In *Proceedings of LRECâĂŹ04*, 2004.

[23] M. Georgescul, A. Clark, and S. Armstrong. Word distributions for thematic segmentation in a support vector machine approach. In *Proceedings of CoNLL*, pages 101–108, 2006.

[24] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, 1992.

[25] H. P. Grice. Utterer's meaning and intentions. *Philosophical Review*, 1969.

[26] H. P. Grice. *Logic and conversation*, page 41âĂŞ58. New York: Academic Press, 1975.

[27] J. Grimes. *The thread of discourse*. The Hague, 1975.

[28] B. Grosz and J. Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, 1992.

[29] B. Grosz and S. Kraus. Collaborative plans for group activities. In *Proceedings of IJCAI-93*, pages 367–373, Chambery, France, 1993.

[30] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 1986.

[31] A. Gruenstein, J. Niekrasz, and M. Purver. Meeting structure annotation: Data and tools. In *Proc. of the SIGdial Workshop on Discourse and Dialogue*, 2005.

[32] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals. Transcription of conference room meetings: An investigation. In *Proc. of Interspeech 2005*, 2005.

[33] M. A. K. Halliday and R. Hasan. *Cohesion in English*. London: Longman, 1976.

[34] D. Harman and P. Over. The effects of human variation in duc summarization evaluation. In *Proceedings of the Workshop on Text Summarization Branches Out of ACL 2004*, 2004.

[35] M. Hearst. TextTiling: Segmenting text into multiparagraph subtopic passages. *Computational Linguistics*, 25(3):527–571, 1997.

[36] J. Hirschberg and C. H. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. of ACL 1996*, 1996.

[37] J. R. Hobbs. Coherence and coreference. *Cognitive Science*, 3:67ï£¡V90, 1979.

[38] J. R. Hobbs. Abduction in natural language understanding. In L. Horn and G. Ward, editors, *Handbook of Pragmatics*. Blackwell, 2004.

[39] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, 1999.

[40] P. Hsueh and J. Moore. Automatic topic segmentation and lablelling in multiparty dialogue. In *the first IEEE/ACM workshop on Spoken Language Technology (SLT) 2006*, 2006.

[41] P. Hsueh, J. Moore, and S. Renals. Automatic segmentation of multiparty dialogue. In *Proc. of EACL 2006*, 2006.

[42] P. Hsueh and J. D. Moore. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the 45th Annual Meeting of the ACL*, 2007.

[43] B. Hutchinson. Acquiring the meaning of discourse markers. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, 2004.

[44] A. Janin et al. The ICSI meeting corpus. In *Proc. of ICASSP 2003*, 2003.

[45] A. Kehler. *Coherence, Reference and the Theory of Gramma*, chapter A Theory of Discourse Coherence. CSLI Publications, Stanford, CA, 2002.

[46] W. Kraaij, A. Smeaton, P.Over, and J. Arlandis. Trecvid 2004 ï£¡ an introduction. In *TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[47] S. C. Levinson. *Pragmatics*. Cambridge University Press, 1983.

[48] G. Levow. Prosody-based topic segmentation for mandarin broadcast news. In *Proc. of HLT 2004*, 2004.

[49] D. Litman and R. Passoneau. Combining multiple knowledge sources for discourse segmentation. In *Proc. of the ACL 1995*, 1995.

[50] W. Mann and S. Thompson. *Rhetorical structure theory: Toward a functional theory of text organization.* 1988.

[51] S. Marchand-Mailet. Meeting record modeling for enhanced browsing. Technical report, Computer Vision and Multimedia Lab, Computer Centre, University of Geneva, Switzerland, 2003.

[52] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):305–317, 2005.

[53] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 1991.

[54] R. Passonneau and D. Litman. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proc. of ACL 1993*, 1993.

[55] F. C. N. Pereira and B. J. Grosz. *Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1994.

[56] L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[57] L. Polanyi. A formal model of discourse structure. *Journal of Pragmatics*, pages 601–638, 1988.

[58] J. Ponte and W. Croft. Text segmentation by topic. In *Proc. of the Conference on Research and Advanced Technology for Digital Libraries 1997*, 1997.

[59] M. Purver, P. Ehlen, and J. Niekrasz. Shallow discourse structure for action item detection. In *the Workshop of HLT-NAACL: Analyzing Conversations in Text and Speech*. ACM Press, 2006.

[60] J. Reynar. *Topic Segmentation: Algorithms and Applications*. PhD thesis, UPenn, PA USA, 1998.

[61] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696âĂŞ735, 1974.

[62] J. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University, Cambridge England, 1969.

[63] J. R. Searle. *Collective intentionality*. 1990.

[64] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communications*, 31(1-2):127–254, 2000.

[65] N. Stokes, J. Carthy, and A. Smeaton. Select: a lexical cohesion based news story segmentation system. *AI Communications*, 17(1):3–12, Jan. 2004.

[66] TDT-Evaluation. The 2002 topic detection and tracking (tdt2002) task definition and evaluation plan. Technical report, TOPIC DETECTION AND TRACKING (TDT2002), 2002.

[67] D. Trieschnigg and W. Kraaij. Hierarchical topic detection in large digital news archives: Exploring a sample based approach. *Journal of Digital Information Management*, 3(1), 2005.

[68] G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57, 2001.

[69] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of the 28th Annual Meeting of the ACL*, 2001.

[70] P. van Mulbregt, J. Carp, L. Gillick, S. Lowe, and J. Yamron. Segmentation of automatically transcribed broadcast news text. In *Proceedings of the DARPA Broadcast News Workshop*, pages 77–80. Morgan Kaufman Publishers, 1999.

[71] F. Wolf and E. Gibson. Representing discourse coherence: a corpus-based analysis. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 134, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[72] K. Zechner and A. Waibel. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proc. of COLING-2000*, 2000.

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

Technologies for Automatic Summarization

November 6, 2007

AMI Consortium State of the Art Report

Technologies for Automatic Summarization

November 6, 2007

Abstract

In this document we provide a taxonomy of approaches to automatic summarization and a historical overview of both text and speech summarization. For speech summarization, we focus primarily on four popular domains of research: broadcast news, meetings, lectures and voicemail. The purpose of this document is primarily to place current speech summarization in the proper historical context, and to increase links between text summarization and speech summarization researchers.

1 Types of Summaries

One possible division of automatic summaries is between *extracts* and *abstracts*, where the former consists of units removed from the source text and concatenated together in a new, shorter document, and the latter consists of novel sentences representing the source document from a more high-level perspective. Rather than being a hard division, however, abstracts and extracts exist on a single continuum, and extracts can potentially be made more abstract-like through further interpretation or transformation of the data. Simple extracts can also be more than simply cutting and pasting; the extracted units can be compressed, made less disfluent, ordered to maximize coherence, and merged to reduce redundancy, to give a few examples.

Another possible division of summaries is between *indicative* and *informative* summaries. An *informative* summary is meant to convey the most important information of the source text, thus acting as a substitute for the original text. On the other hand, an *indicative* summary acts as a guide for where to find the most important parts of the source text. Using these definitions, the summaries we are creating in this current research can serve as either type depending on the use case. The summaries are incorporated into a meeting browser, and a time-constrained user can either read the summary in place of the entire transcript and/or use the summary as an efficient way of indexing into the meeting record.

Another division is between *multiple-document* and *single-document* summaries. In the latter case, information is gleaned from several source documents and summarized in a single output document; in these cases, redundancy is much more of an issue than with single-document summarization. In this research, we focus on summaries of individual meetings, but many of the methods are easily extendable to the task of summarizing and linking multiple archived meetings.

Similarly, this work focuses on *generic* summaries rather than *query-dependant* summaries, but the methods could be extended to query-dependent summarization. In generic summarization, each summary is created without regard to any specific information need,

based on the inherent informativeness of the document. For query-dependent summarization, units are extracted based partly on how similar they are to a user-supplied query or information need.

It is possible to divide between *text* and *speech* summarization, or *text* and *multi-media* summarization, in the sense that the fields of research have separate but overlapping histories and use different types of data as input (and potentially as output as well), but of course the simplest way to approach speech summarization is to treat is as a text summarization problem, using a noisy text source. Speech summarization and text summarization approaches often use many of the same features or types of features. However, a central thesis of this work is that it is advantageous to use speech-specific features at various steps of the summarization process, compared with simply treating the problem as a text summarization task.

2 Previous Work

2.1 Text Summarization

Among the earliest work on automatic text summarization is the research by Luhn [25], who particularly focused on recognizing keywords in text. Luhn was among the first to recognize that the words with highest resolving power are words with medium or moderately high frequency in a given document.

A decade later, Edmundson [6] began to look beyond keywords for the summarization of scientific articles. He focused on four particular areas of interest: cue phrases, keywords, title words, and location. While keyword detection had been the subject of previous research the other areas were novel. Cue phrases are phrases that are very likely to signal an important sentence, and could include phrases such as “significantly”, “in conclusion” or “impossible” in the scientific articles domain. On the other hand, there are so-called Stigma phrases that may signal unimportance: specifically, these might be hedging or belittling expressions. Also particular to the type of academic articles Edmundson was working with is the Title feature, which weights each sentence according to how many times its constituent words occur in section or article titles. And finally, the Location feature weights sentences more highly if they occur under a section heading or occur very early or late in the article. Edmundson’s summarization system then works by scoring and extracting sentences based on a linear combination of these four features. These categories of features are still used today, though more often in machine-learning frameworks.

The ADAM system of the 1970s [30] relied heavily on cue phrases, but also strove to maximize coherence by analyzing whether a candidate sentence contained anaphoric references [7]. In the case that a candidate did, the system tried to either extract the preceding sentences as well or to re-write the candidate sentence so that it could stand alone. If neither of these were possible, the candidate was not chosen.

In the 1980s, several summarization methods arose that were inspired by findings in psychology and cognitive science [5, 9, 18]. These methods generally use human processing and understanding of text as a model for automatic abstraction. The source is

interpreted and inferences are made based on prior knowledge. For an automatic summarization method, a schemata might be created relating to the domain of the data being summarized. What differentiates these methods from the earlier summarization methods described above is that the input is *interpreted* and *represented* more deeply than before. For example, the FRUMP system [5] uses "sketchy scripts" to model events in the real-world for the purpose of summarizing news articles. One example would be a sketchy script relating to earthquakes. We have prior knowledge about earthquakes, such as the magnitude on the Richter scale, the location of the epicenter, the number of deaths and the amount of damage inflicted. When a particular sketchy script is activated, these pieces of information are sought in the source data. An interesting overview of such approaches can be found in [7].

Summarization research underwent a major resurgence in the late 1980s and 1990s, primarily due to the explosion of data available from sources such as the web and newswire services. Because of the volume and variety of data to be summarized, the summarization techniques were more often extractive than abstractive, as the former is domain-independent,requires little or no prior knowledge, and can process a large amount of data efficiently. The field therefore tended to move away from the schema-based, cognition-inspired approaches of the 1980s.

Much of the work of this period revisited the seminal work of Edmundson [6] and his investigation of cue phrases, keywords, title words, and location features. The newer work incorporated these same features into machine-learning frameworks where classifiers are trained on human gold-standard extracts [22, 43], rather than manually tuning the weights of these features as Edmundson did. For the tasks of summarizing engineering papers [22] and computational linguistics papers [43], the most useful features were found to be cue phrases and locational features.

During this same period, other researchers investigated the use of rhetorical relations for the purpose of text summarization, particularly in the framework of Rhetorical Structure Theory (RST) [26]. A hypothesis of RST is that a given document can be represented as a single binary-branching rhetorical tree comprised of nuclei-satellite pairs, where a particular rhetorical relation exists between each nuclei-satellite pair. By pruning such a rhetorical tree, a summary of the entire text can be generated [37, 27, 28].

Contemporary work utilized linguistics resources such as WordNet, a database of lexical semantics, in order to derive relations between terms or phrases in a document. In work by Barzilay and Elhadad [1] lexical chains were detected according to the relatedness of document terms, and sentences corresponding to the strongest chains were extracted. The SUMMARIST system [15] utilized WordNet for concept detection in the summarization of news articles.

Also in the late 1990s, interest in multi-document summarization was growing. Creating a single summary of multiple documents presented, and still presents, and interesting challenge, as the summarizer must determine which documents are relevant to a given query and/or related to one another and must not extract the same information from multiple sources. In other words, the problem of *redundancy* is paramount. Carbonell and Goldstein [2] introducted the Maximal Marginal Relevance (MMR) algorithm, which scores a candidate sentence according to how relevant it is to a query (or how generally rele-

vant, for a generic summary) and how similar it is to sentences that have already been extracted. The latter scores is used to penalize the former, thereby reducing redundancy in the resultant summary. MMR remains popular both as a stand-alone algorithm in its own right as well as a feature score in more complex summarization methods. Work by Radev et. al [39, 38] addressed single- and multi-document summarization via a centroid-method. A centroid is a pseudo-document consisting of important terms and their associated term-weight scores, representing the source document(s) as a whole. The authors address the redundancy probelem via the idea of cross-sentence information subsumption, whereby sentences that are too similar to other sentences are penalized, similar to the MMR method.

The work of Maybury [31] extended summarization work from merely processing and summarizing text to summarizing multi-modal event data. In the domain of battle simulation, the researchers took as input battle events such as missile fire, refueling, radar sweeps and movement and generated summaries based on the frequencies of such events and relations between such events. Not only are the inputs multi-modal events, but the output can be a combination of textual and graphical summaries in order to expedite perception and comprehension of the battle scene. The researchers also take into account that such summaries should be tailored to the user: for example, an intelligence officer might care more about enemy size and position whereas a logistician will care about refueling and supplies.

Since 2001, the Document Understanding Conference [1] has encouraged research in the area of multi-document, query-dependent summarization. For the text summarization community, this annual conference provides the benchmark tasks for comparing and evaluating state-of-the-art summarization systems. While the data used has primarily been newswire data, DUC has recently added tracks relating to the summarization of weblog opinions. Though a wide variety of systems have been entered in DUC, one finding is that the most competitive systems have extensive query-expansion modules. In fact, query-expansion forms the core of many of the systems [23, 16].

2.2 Speech Summarization

Chen and Withgott [3] identified areas of emphasis in speech data in order to create audio summaries, reporting results on two types of data: a recorded interview and telephone speech. The emphasis detection was carried out by training a hidden markov model on training data in which words had been manually labelled for varying degrees of emphasis. The features used in the model were purely prosodic, namely F0 and energy features. The authors reported near-human performance in selecting informative excerpts.

Rohlicek et. al [41] created brief summaries, or gists, of conversations in the air-traffic control domain. The basic summarization goals were to identify flight numbers and classify the type of flight, e.g. *takeoff* or *landing*. Such a system required components of speaker segmentation, speech recognition, natural language parsing and topic classification. The authors reported that the system achieved 98% precision of flight classification with 68% recall.

1. http://duc.nist.gov

One of the early projects on *speech* summarization was VERBMOBIL [40], a speech-to-speech translation system for the domain of travel planning. The system was capable of translating between English, Japanese and German. Though the focus of the project was on speech-to-speech translation, an abstractive summarization facility was added that exploited the information present in the translation module's knowledge sources. A user could therefore be provided with a summary of the dialogue, so that they can confirm the main points of the dialogue were translated correctly, for example. The fact that VERBMOBIL was able to incorporate abstractive summarization is due to the fact that the speech was limited to a very narrrow domain of travel planning and hotel reservation; normally it would be very difficult to create such structured abstracts in unrestricted domains.

Simultaneously work was being carried out on the MIMI dialogue summarizer [20], which was used for the summarization of spontaneous conversations in Japanese. Like VERBMOBIL, these dialogues were in a limited domain; in this case, negotiations for booking meetings rooms. The system creates a running transcript of the transactions so far, by recognizing domain-specific patterns and merging redundant information.

2.2.1 Summarization of Newscasts

One of the domains of speech summarization that has received the most attention and perhaps has the longest history is the domain of broadcast news summarization. Summarizing broadcast news is an interesting task, as the data consists of both spontaneous and read segments and so represents a middle-ground between text and spontaneous speech summarization. In Hirschberg et. al [12], a user interface tool is provided for browsing and information retrieval of spoken audio - in this case, National Public Radio broadcasts. The browser adds audio paragraphs, or *paratones*, to the speech transcript, using intonational information. This is a good example of how structure can be added to unstructured speech data in order make it more readable as well as more amenable to subsequent analysis incorporating structural features. Their browser also highlights keywords in the transcript based on acoustic and lexical information.

In Valenza et. al [44], summarization of the American Broadcast News corpus is carried out by weighting terms according to an acoustic confidence measure and a term-weighting metric from information retrieval called inverse frequency (described in detail in a later chapter). The units of extraction are n-grams, utterances and keywords, which are scored according to the normalized sums of their constituent words in the case of n-grams and utterances. When a user desires a low word-error rate (WER) above all else, a weighting parameter can be changed to favor the acoustic confidence score over the lexical score. One of the most interesting results of this work is that the WER of summaries portions are typically much lower than the overall WER of the source data, a finding that has since been attested in other work [33]. Valenza et. al also provide a simple but intuitive interface for browing the recognizer output.

In work by Hori and Furui [13] on Japanese broadcast news summarization, each sentence has a subset of its words extracted based on each word's topic score – a measure of its significance – and a concatenation likelihood, the likelihood of the word being concatenated to the previously extracted segment. Using this method, they report that 86% of the

important words in the test set are extracted.

More recently in the broadcast news domain, Maskey and Hirschberg [29] found that the best summarization results utilized prosodic, lexical and structural features, but that prosodic features alone resulted in good-quality summarization. The prosodic features they investigated were broadly features of pitch, energy, speaking rate and sentence duration. Work by Ohtake et. al [36] explored using *only* prosodic features for speech-to-speech summarization of Japanese newscasts, finding that such summaries rated comparably with a system relying on speech recognition output.

2.2.2 Summarization of Meetings

In the domain of meetings, Waibel et. al [45] implemented a modified version of maximal marginal relevance applied to speech transcripts, presenting the user with the *n* best sentences in a meeting browser interface. The browser contained several information streams for efficient meeting access, such as topic-tracking, speaker activity, audio/video recordings and automatically-generated summaries. However, the authors did not research any speech-specific information for summarization; this work was purely text summarization applied to speech transcripts.

Zechner [46] investigated summarizing several genres of speech, including spontaneous meeting speech. Though relevance detection in his work relied largely on *tf.idf* scores, Zechner also explored cross-speaker information linking and question/answer detection, so that utterances could be extracted not only according to high *tf.idf* scores, but also if they were linked to other informative utterances.

On the ICSI corpus, Galley [10] used skip-chain Conditional Random Fields to model pragmatic dependencies such as QUESTION-ANSWER between paired meeting utterances, and used a combination of lexical, prosodic, structural and discourse features to rank utterances by importance. The types of features used were classified as *lexical features*, *information retrieval features*, *acoustic features*, *structural and durational featuers* and *discourse features*. Galley found that while the most useful single feature class was *lexical* features, a combination of acoustic, durational and structural features exhibited comparable performance according to Pyramid evaluation.

Simpson and Gotoh [42], also working with the ICSI meeting corpus, investigated speaker-independent prosodic features for meeting summarization. A problem of working with features relying on absolute measurements of pitch and energy is that these features vary greatly depending on the speaker and the meeting conditions, and thus require normalization. The authors therefore investigated the usefulness of speaker-indepedent features such as pauses, pitch and energy changes across pauses, and pitch and energy changes across units. They found that pause durations and pitch changes across units were the most consistent features across multiple speakers and multiple meetings.

Liu et. al [24] reported the results of a pilot study on the the effect of disfluencies on automatic speech summarization, using the ICSI corpus. They found that the manual removal of disfluencies did not improve summarization performance according to the ROUGE metric.

In our own work on the ICSI corpus, Murray et al. [33, 34] compared text summarization approaches with feature-based approaches incorporating prosodic features, with human judges favoring the feature-based approaches. In subsequent work [35], we began to look at additional speech-specific characteristics such as speaker and discourse features. One significant finding of these papers was that the ROUGE evaluation metric did not correlate well with human judgments on this test data.

2.2.3 Summarization of Lectures

Hori et al. [14] have developed an integrated speech summarization approach, based on finite state transducers, in which the recognition and summarization components are composed into a single finite state transducer, reporting results on a lecture summarization task.

Also in the lectures domain, Fujii et. al [8] attempted to label cue phrases and use cue phrase features in order to supplement lexical and prosodic features in extractive summarization. They reported that the use of cue phrases for summarization improved the summaries according to both f-scores and ROUGE scores.

Zhang et. al [47] compared feature types for summarization across domains, concentrating on lecture speech and broadcast news speech in Mandarin. They found that acoustic and structural features are more important for broadcast news than for the lecture task, and that the quality of broadcast news summaries is less dependent on ASR performance.

2.2.4 Voicemail Summarization

The SCANMail system [11] was developed to allow a user to navigate their voicemail messages in a graphical user interface. The system incorporated information retrieval and information extraction components, allowing a user to query the voicemail messages, and automatically extracting relevant information such as phone numbers. Huang et. al [17] and Jansche and Abbey [19] also described techniques for extracting phone numbers from voicemails.

Koumpis and Renals [21] investigated prosodic features for summarizing voicemail messages in order to send voicemail summaries to mobile devices. They reported that while the optimal feature subset for classification was the lexical subset, an advantage could be had by augmenting those lexical features with prosodic features, especially pitch range and pause information.

2.3 From Text to Speech

McKeown et. al [32] provided an overview of text summarization approaches and discussed how text-based methods might be extended to speech data. The authors described the challenges in summarizing differing speech genres such as Broadcast News and meeting speech and which features are useful in each of those domains. Their summarization work involved components of speaker segmentation, topic segmentation, detection of agreement/disagreement, and prosodic modelling, among others.

Christensen et. al [4] investigated how well text summarization techniques for newswire data could be extended to broadcast news summarization. In analyzing feature subsets, they found that positional features were more useful for text summarization than for broadcast news summarization and that positional features alone provided very good results for text. In contrast, no single feature set in their speech summarization experiments was as dominant, and all of the features involving position, length, term-weights and named entities made significant contributions to classification. They also found that increased word-error rate only caused slight degradation according to their automatic metrics, but that human judges rated the error-filled summaries much more severely.

3 Conclusion

In this document we have provided an overview of summarization types and a literature review of both text and speech summarization, looking particularly at speech domains of broadcast news, meetings and lectures. While there are certainly other interesting speech genres, speech summarization research has been focused largely on these few domains to date. It is hoped that by reviewing text summarization and speech summarization together, the best ideas of one community can inform the other and increase links between the parallel fields of research.

References

[1] R. Barzilay and M. Elhadad. Using lexical chains for summarisation. In *Proc. of ACL 1997, Madrid, Spain*, pages 10–18, 1997.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval 1998, Melbourne, Australia*, pages 335–336, 1998.

[3] F. Chen and M. Withgott. The use of emphasis to automatically summarize a spoken discourse. In *Proc. of ICASSP 1992, San Francisco, USA*, pages 229–232, 1992.

[4] H. Christensen, Y. Gotoh, B. Kolluru, and S. Renals. Are extractive text summarisation techniques portable to broadcast news? In *Proc. of IEEE Speech Recognition and Understanding Workshop, St. Thomas, USVI*, pages 489–494, 2003.

[5] G. DeJong. An overview of the FRUMP system. In W. G. Lehnert and M. H. Ringle, editors, *Strategies for Natural Language Processing*, pages 149–176. Lawrence Erlbaum Associates, Publishers, 1982.

[6] H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2):264–285, 1969.

[7] B. Endres-Niggemeyer. *Summarizing Information*. Springer, Berlin, 1998.

[8] Y. Fujii, N. Kitaoka, and S. Nakagawa. Automatic extraction of cue phrases for important sentences in lecture speech and automatic lecture speech summarization. In *Proc. of Interspeech 2007, Antwerp, Belgium*, pages 2801–2804, 2007.

[9] D. Fum, G. Guida, and C. Tasso. Forward and backward reasoning in automatic abstracting. In *Proc. of the (COLING '82), Prague, Czech Republic*, pages 83–88, 1982.

[10] M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of EMNLP 2006, Sydney, Australia*, pages 364–372, 2006.

[11] J. Hirschberg, M. Bacchiani, D. Hindle, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, S. Whittaker, and G. Zamchick. Scanmail: Browsing and searching speech data by content. In *Proc. of Interspeech 2001, Aalborg, Denmark*, pages 1299–1302, 2001.

[12] J. Hirschberg, S. Whittaker, D. Hindle, F. Pereira, and A. Singhal. Finding information in audio: A new paradigm for audio browsing and retrieval, 1999.

[13] C. Hori and S. Furui. Automatic speech summarization based on word significance and linguistic likelihood. In *Proc. of ICASSP 2000, Istanbul, Turkey*, pages 1579–1582, 2000.

[14] T. Hori, C. Hori, and Y. Minami. Speech summarization using weighted finite-state transducers. In *Proc. of Interspeech 2003, Geneva, Switzerland*, pages 2817–2820, 2003.

[15] E. Hovy and C.-Y. Lin. Automated text summarization in summarist. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94. MITP, 1999.

[16] E. Hovy, C.-Y. Lin, and L. Zhou. A BE-based multi-document summarizer with query interpretation. In *Proc. of DUC 2005, Vancouver, CA*, 2005.

[17] J. Huang, G. Zweig, and M. Padmanabhan. Information extraction from voicemail. In *Proc. of ACL 2001, Toulouse, France*, pages 290–297, 2001.

[18] P. Jacobs and L. Rau. SCISOR: Extracting information from on-line news. *CACM*, 33(11):88–97, 1990.

[19] M. Jansche and S. Abney. Information extraction from voicemail transcripts. In *Proc. of EMNLP 2002, Philadelphia, USA*, pages 320–327, 2002.

[20] M. Kameyama and I. Arima. Coping with aboutness complexity in information extraction from spoken dialogues. pages 87–90, 1994.

[21] K. Koumpis and S. Renals. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Transactions on Speech and Language Processing*, 2:1–24, 2005.

[22] J. Kupiec, J. Pederson, and F. Chen. A trainable document summarizer. In *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA*, pages 68–73, 1995.

[23] F. Lacatusu, A. Hickl, P. Aarseth, and L. Taylor. Lite-GISTexter at DUC 2005. In *Proc. of DUC 2005, Vancouver, CA*, 2005.

[24] Y. Liu, F. Liu, B. Li, and S. Xie. Do disfluencies affect meeting summarization: A pilot study on the impact of disfluencies. In *Proc. of MLMI 2007, Brno, Czech Republic*, page poster, 2007.

[25] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165, 1958.

[26] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[27] D. Marcu. Discourse trees are good indicators of importance in text. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 123–136. MITP, Cambridge, MA, 1995.

[28] D. Marcu. From discourse structures to text summaries. In *Proc. of ACL 1997*

Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, pages 82–88, 1997.

[29] S. Maskey and J. Hirschberg. Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 621–624, 2005.

[30] B. Mathis. *Techniques for the evaluation and improvement of computer-produced abstracts.* Ohio State University Technical Report OSU-CISRC-TR-72-15, Ohio, USA, 1972.

[31] M. Maybury. Generating summaries from event data. *IPM*, 31(5):735–751, September 1995.

[32] K. McKeown, J. Hirschberg, M. Galley, and S. Maskey. From text to speech summarization. In *Proc. of ICASSP 2005, Philadelphia, USA*, pages 997–1000, 2005.

[33] G. Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 593–596, 2005.

[34] G. Murray, S. Renals, J. Carletta, and J. Moore. Evaluating automatic summaries of meeting recordings. In *Proc. of the ACL 2005 MTSE Workshop, Ann Arbor, MI, USA*, pages 33–40, 2005.

[35] G. Murray, S. Renals, J. Moore, and J. Carletta. Incorporating speaker and discourse features into speech summarization. In *Proc. of the HLT-NAACL 2006, New York City, USA*, pages 367–374, 2006.

[36] K. Ohtake, K. Yamamoto, Y. Toma, S. Sado, S. Masuyama, and S. Nakagawa. Newscast speech summarization via sentence shortening based on prosodic features. In *Proc. of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan*, pages 167–170, 2003.

[37] K. Ono, K. Sumita, and S. Miike. Abstract generation based on rhetorical structure extraction. In *Proc. of COLING 1994, Kyoto, Japan*, pages 344–348, 1994.

[38] D. Radev, S. Blair-Goldensohn, and Z. Zhang. Experiments in single and multidocument summarization using mead. In *Proc. of DUC 2001, New Orleans, LA, USA*, 2001.

[39] D. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *Proc. of the ANLP/NAACL 2000 Workshop, Seattle, WA*, pages 21–29, 2000.

[40] N. Reithinger, M. Kipp, R. Engel, and J. Alexandersson. Summarizing multilingual spoken negotiation dialogues. In *Proc. of ACL 2000, Hong Kong*, pages 310–317, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[41] J. R. Rohlicek. Gisting continuous speech. In *Proc. of ICASSP 1992, San Francisco, USA*, pages 384–384, 1992.

[42] S. Simpson and Y. Gotoh. Towards speaker independent features for information extraction from meeting audio data. In *Proc. of MLMI 2005, Edinburgh, UK*, page poster, 2005.

[43] S. Teufel and M. Moens. Sentence extraction as a classification task. In *Proc. of ACL 1997, Workshop on Intelligent and Scalable Text Summarization, Madrid, Spain*, pages 58–65, 1997.

[44] R. Valenza, T. Robinson, M. Hickey, and R. Tucker. Summarization of spoken audio through information extraction. In *Proc. of the ESCA Workshop on Accessing Information in Spoken Audio, Cambridge UK*, pages 111–116, 1999.

[45] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In D. E. M. Penrose, editor, *Proc. of the Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, USA*, pages 281–286, 1998.

[46] K. Zechner. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485, 2002.

[47] J. Zhang, H. Chan, P. Fung, and L. Cao. Comparative study on speech summarization of broadcast news and lecture speech. In *Proc. of Interspeech 2007, Antwerp, Belgium*, pages 2781–2784, 2007.

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

Automatic Dialogue Act Recognition

November 6, 2007

AMI Consortium State of the Art Report

Automatic Dialogue Act Recognition

November 6, 2007

Abstract

A DA is a construct that describes the role that an utterance plays in a conversation and provides a bridge between an orthographic (word-level) transcription, and a richer representation of the discourse. The reliable recognition of the DA sequence in a conversation, and the resulting knowledge of the discourse structure, plays an important role in the development of applications such as: action items detection, decision detection, automatic summarisation, topic segmentation, dialogue structure annotation, etc. DA recognition systems are usually based on supervised statistical approaches: a model is learned (trained) from a set of annotated examples, and then evaluated on unseen data. This process requires to collect and manually annotate relevant amounts of conversational data. Therefore several annotated corpora have been produced in the last decade, giving birth to multiple DA annotation schemes. The DA recognition task comprises two related sub-tasks: segmentation, and classification or tagging. DA segmentation consists of subdividing the conversation into unlabelled DA segments closer to those manually annotated. Unlabelled DA segments are then classified and tagged with the most likely DA label. These tasks may be performed concurrently (joint DA recognition) or sequentially. Multiple evaluation metrics have been proposed for segmentation, classification and the DA recognition task.

1 Introduction

The concept of dialogue acts (DAs) is based on the speech acts described by Austin [1962] and by Searle [1969]. The idea is that speaking is acting on several levels, from the mere production of sound, over the expression of propositional content to the expression of the speaker's intention and the desired influence on the listener. Dialogue acts are labels for utterances which roughly categorise the speaker's intention.

As such, they are useful for various purposes in a dialogue or meeting processing situation. For example DAs can be used as elements in a structural model of a meeting. A simple example would be a browser which highlights all points where a suggestion or offer was recognised. Often DA labels serve also as elementary units to recognise higher levels of structure in a discourse. DAs may also control the processing of discourse content. To generate abstractive summaries, for example, content is extracted from utterances, and integrated in a discourse memory depending on the DAs of the utterances.

The dialogue act recognition process consists of two subtasks: segmentation and classification (tagging). The first step is to subdivide the sequence of transcribed words in terms of DA segments. The goal is to segment the text into utterances that have approximately similar temporal boundaries to the annotated DA units. The second step is to classify each segment as one of the DA classes from the adopted DA annotation scheme. These two steps may be performed either sequentially (segmentation followed by classification) or

jointly (both tasks carried out simultaneously by an integrated system). Although most of the work on automatic DA processing have been focused on the tagging task, assuming knowledge of the reference DA segmentation; novel integrated DA recognition frameworks are growing in popularity.

2 Dialogue Act Annotated Data Resources

Any effort to recognize dialogue acts requires data. The usual practice is to employ supervised machine learning, using material that has been hand-transcribed and then hand-annotated with a suitable dialogue act scheme. Since creating this sort of data is expensive, most efforts re-use an existing data set, or corpus, wherever they can. There are a number of factors that need to be balanced in deciding on a corpus:

- how much data is available, since more data usually implies better results.
- how many dialogue act classes the scheme contains. Classifiers have trouble learning too many distinctions.
- how well distributed the classes are. Classifiers work best with relatively equal numbers of examples for the various classes.
- how reliable the hand-annotation is. If there are several human annotators involved and they tend to assign different classes for the same kind of material, the inconsistency makes it difficult both for the classifier and for evaluating the results. On the other hand, if there is only one annotator, but no one else would assign classes the same way, what the classifier is learning may not be useful.
- what language the data is for . The vast majority of available material and published work is on English.
- whether the data is generally available or access is restricted in some way.

For researchers who are interested in dialogue act recognition as an end in itself, for instance, as a means of trying out various machine learning algorithms, these are the primary considerations. However, where the recognizer is being built for use in an end application, it is also important that the dialogue act scheme makes the distinctions that the end application actually needs. It is no use having an extremely accurate classifier that cannot identify "backchannel" utterances such as "mhm-hmm" in a system that requires a very natural style of interaction, for instance. In addition, it is important that the material to which a scheme has been applied be similar enough to what the end application will encounter for what the classifier learns to transfer well. The closer the material, the better, which is why systems developers almost always collect at least some human-human dialogues that are as close to what the system will do as they can get. Since most applications involve having a system perform a task for the user, such as booking travel, task-oriented data is of the most use, but simply making the distinction between task-oriented dialogues and more free-ranging conversations is not enough. Differences such as using non-native or elderly speakers can have a large effect where these are the target users for the end application. Similarly, whether or not speakers use telephones changes their behaviour.

In dialogue act recognition, it is not necessary to learn every label from the hand-annotated scheme. Hand-annotated data can be transformed into a smaller set of labels by grouping individual classes together. This sort of transformation is sometimes called a "classmap".

Because of this practice, recent dialogue act scheme designers often include more labels than a classifier can learn, and then perform an analysis of the hand-annotated data once it is complete in order to decide what transform to use. There can be several acceptable classmaps for the same scheme, depending on how the resulting classifier is to be used. In constructing classmaps, it is common to put together classes that the human annotators frequently confuse with each other to make the data more consistent. However, it is important to ensure the label groupings also make sense in terms of the end application. The smallest schemes tend to provide 12-15 mutually exclusive labels and expect only a few, or none, to be combined.

Dialogue act schemes also differ in how they instruct the human annotator to segment the dialogue. For some schemes, the segmentation is purely ideational; annotators are to decide on segmentation by breaking the material into pieces that each express a complete meaning. For others, the segmentation is partly mechanical – for instance, the annotator may be instructed to provide segment boundaries at long pauses. Occasionally, the scheme assumes that the material has been presegmented by completely mechanical means (e.g., as in the Japanese Map Task Corpus, [Horiuchi et al., 1999]). Not surprisingly, ideational segmentations show the most disagreement among the human annotators, but they also in theory supply the most information.

Finally, dialogue act schemes vary in whether they adhere to dialogue act theory in simply segmenting and labelling material based on speaker intentions, or whether they include labels that are, strictly speaking, not dialogue acts at all. The most common addition is labels that identify disfluent material, particularly at the beginning of turns. The reason for their inclusion is to improve the results of language modelling on the data. Theoretically, the disfluent material belongs within an adjacent act, but because questions typically have a different syntactic form from statements and commands, the words that occur at the beginning of an act are important for determining what the act is. Dialogue act schemes that include these quasi-acts assume they will be used to strip this material out, sometimes as a first step before proper dialogue act recognition.

Klein et al. [1998] provides a detailed survey of early dialogue act schemes. The corpora currently in most common use for dialogue act recognition are the following:

The HCRC Map Task Corpus [Anderson et al., 1991], using a scheme developed for it [Carletta et al., 1997]. The scheme is intended to be general, but the material coded involves two people navigating around a simple map. One unusual aspect of this material is that it contains higher level dialogue structure coding. Many corpus users consider the fact that the speakers are Scottish as a disadvantage. It is also relatively small: 128 dialogues resulting in around 10 hours of speech.

The related DCIEM Map Task Corpus [Bard et al., 1996], which replicates the same task but using Canadian army reservists and includes sleep deprivation conditions comparing the effects of various drugs.

The Switchboard Corpus [Godfrey et al., 1992] using SWBD-DAMSL [Jurafsky et al., 1997b]. The underlying material consists of telephone conversations on a fixed set of topics, resulting in more than 200000 utterances and 1.4 millions transcribed words. The SWBD-DAMSL annotation scheme comprises 226 unique tags, which were subsequently clustered into 42 broad DA classes. A common concern when using this material is that that a very large proportion of the dialogue acts are of the

same type (basic statements).

The ICSI Meeting Corpus [Janin et al., 2003], using the ICSI-MRDA scheme [Shriberg et al., April-May 2004]. The corpus contains audio recordings of research group meetings. The scheme requires each act to be labelled along a number of semi-orthogonal dimensions, with thousands of tag combinations that are theoretically possible.

The AMI Meeting Corpus [Carletta, In Press], using a scheme developed for it. This corpus is unusual in involving non-native speakers of English and in making available a range of videos and other outputs that capture behaviour more fully than usual. The dialogue act scheme includes some extra features related to acts, such as information about addressing and some very rudimentary discourse structure.

2.1 The ICSI Meeting Corpus and Dialogue Act Tag Set

The ICSI meetings corpus [Janin et al., 2003] consists of 75 naturally occurring research group meetings at the International Computer Science Institute in Berkeley during the years 2000–2002, and recorded using close-talking microphones worn by each participant (in addition, there were also four tabletop microphones). Each meeting lasts about one hour and involves an average of six participants, resulting in about 72 hours of multichannel audio data. The corpus contains human-to-human interactions recorded from naturally occurring meetings. Moreover, having different meeting topics and meeting types, the data set is heterogeneous both in terms of content and structure.

Orthographic transcriptions are available for the entire corpus, and each meeting has been manually segmented and annotated in terms of Dialogue Acts, using the ICSI MRDA scheme [Shriberg et al., April-May 2004]. The MRDA scheme, outlined in table 1, is based on a hierarchy of DA types and sub-types (11 generic tags and 40 specific sub-tags), and allows multiple sub-categorisations for a single DA unit. A DA is usually composed by a single generic tag (statement, question, etc.) and several specific sub tags. This extremely rich annotation scheme results in more than a thousand unique DAs, although many are observed infrequently. To reduce the number of sparsely observed categories, a reduced set of five broad DA categories has been defined in [Ang et al., 2005, Zimmermann et al., 2006a]. Unique DAs were manually grouped into five generic categories: statements, questions, backchannels, fillers and disruptions. The distribution of these categories across the corpus is shown in table 2. Note that statements are the most frequently occurring unit, and also the longest, having an average length of 2.3 seconds (9 words). All the other categories (except backchannels which usually last only a tenth of a second) share an average length of 1.6 seconds (6 words). An average meeting contains about 1500 DA units.

In order to have directly comparable results a formal subdivision has been proposed by Ang et al. [2005]: a training set of 51 meetings (about 80.000 DAs), 11 meetings for the development task (13.500 DAs), and a test set composed by 11 meetings and 15.000 DAs. This leaves out 2 of the 75 meetings, which were excluded because of their different nature.

Statement	
s	**Statement**

Questions	
qy	**Yes/No Question**
qw	**Wh-Question**
qr	**Or Question**
qrr	**Or Clause After Y/N Question**
qo	**Open-ended Question**
qh	**Rhetorical Question**

Floor Management	
fg	**Floor Grabber**
fh	**Floor Holder**
h	**Hold**

Backchannels	
b	**Backchannel**
bk	Acknowledgement
ba	Assessment/Appreciation
bh	Rhetorical Question Backchannel

Responses	
aa	Accept
aap	Partial Accept
na	Affirmative Answer
ar	Reject
arp	Partial Reject
nd	Dispreferred Answer
ng	Negative Answer
am	Maybe
no	No Knowledge

Action Motivators	
co	Command
cs	Suggestion
cc	Commitment

Checks	
f	Follow Me
br	Repetition Request
bu	Understanding Check

Restated Information	
r	Repeat
m	Mimic
bs	Summary
bc	Correct Misspeaking
bsc	Self-Correct Misspeaking

Supportive Functions	
df	Defending/Explanation
e	Elaboration
2	Collaborative Completion

Politeness Mechanisms	
bd	Downplayer
by	Sympathy
fa	Apology
ft	Thanks
fw	Welcome

Further Descriptions	
fe	Exclamation
t	About-Task
tc	Topic Change
j	Joke
t1	Self Talk
t3	Third Party Talk
d	Declarative Question
g	Tag Question
rt	Rising Tone

Disruptions	
%	*Indecipherable*
%-	*Interrupted*
%–	*Abandoned*
x	*Nonspeech*

Nonlabeled	
z	Nonlabeled

Table 1: DA labels used for the annotation of the ICSI meeting corpus: **generic tags**, specific tags and *disruptions*.

Dialogue Act	% of total DA units	% of corpus length
Statement	58.2	74.5
Disruption	12.9	10.1
Backchannel	12.3	0.9
Filler	10.3	8.7
Question	6.2	5.8

Table 2: Distribution of DAs by % of the total number of DA units and by % of corpus length.

2.2 The AMI Dialogue Act Tag Set

The AMI meeting corpus [Carletta et al., 2005] is a multimodal collection of annotated meeting recordings. It consists of about 100 hours of meetings collected in three instrumented meeting rooms. About two thirds of the corpus consists of meetings elicited using a scenario in which four meeting participants, playing different roles on a team, take a product development project from beginning to completion. The scenario portion of the corpus consists of a number of meeting series, with four meeting per series. Each series of four meetings involves the same four participant roles, and comprises project kick-off, functional design, conceptual design, and detailed design meetings. The aim of the corpus collection was to obtain a multimodal record of the complete communicative interaction between the meeting participants. To this end, the meeting rooms were instrumented with a set of synchronised recording devices, including lapel and headset microphones for each participant, an 8-element circular microphone array, six video cameras (four close-up and two room-view), capture devices for the whiteboard and data projector, and digital pens to capture the handwritten notes of each participant. The corpus has been manually annotated at several levels, including orthographic transcriptions, various linguistic phenomena including head and hand movements, and focus of attention[1]. Most of the scenario data in the AMI corpus, over 100,000 utterances, have been annotated for dialogue acts. The DA annotation scheme for the AMI corpus [2] , outlined in table 3, is based around a categorization tailored for group decision making, and consists of 15 dialogue act types (table 3), which are organised in six major groups:

- Information exchange: giving and eliciting information
- Possible actions: making or eliciting suggestions or offers
- Commenting on the discussion: making or eliciting assessments and comments about understanding
- Social acts: expressing positive or negative feelings towards individuals or the group
- Other: a remainder class for utterances which convey an intention, but do not fit into the four previous categories
- Backchannel, Stall and Fragment: classes for utterances without content, which allow complete segmentation of the material

Each DA unit is assigned to a single class, corresponding to the speaker's intent for the utterance. The distribution of the DA classes, shown in table 3, is rather imbalanced,

1. The annotated corpus is freely available from `http://corpus.amiproject.org`
2. Guidelines for Dialogue Act and Addressee Annotation V1.0, Oct 13, 2005. `http://mmm.idiap.ch/private/ami/annotation/dialogue_acts_manual_1.0.pdf`

Group	Dialogue Act		Frequency	
Segmentation	fra	Fragment	14348	14.0%
	bck	Backchannel	11251	11.0%
	stl	Stall	6933	6.8%
Information	inf	Inform	28891	28.3%
	el.inf	Elicit Inform	3703	3.6%
Actions	sug	Suggest	8114	7.9%
	off	Offer	1288	1.3%
	el.sug	Elicit Offer or Suggestion	602	0.6%
Discussion	ass	Assessment	19020	18.6%
	und	Comment about Understanding	1931	1.9%
	el.ass	Elicit Assessment	1942	1.9%
	el.und	Elicit Comment about Understanding	169	0.2%
Social	be.pos	Be Positive	1936	1.9%
	be.neg	Be Negative	77	0.1%
Other	oth	Other	1993	2.0%
Total			102198	100.0%

Table 3: The AMI Dialogue act scheme, and the DA distribution in the annotated scenario meetings.

with over 60% of DAs corresponding to one of the three most frequent classes (inform, backchannel or assess). Over half the DA classes account for less than 10% of the observed DAs. This annotation scheme is different to the one used for the ICSI corpus (section 2.1), thus it is not possible to test a DA recognition system developed on the AMI data on the ICSI corpus or vice-versa.

The scenario meetings are organised in 35 series of (normally) four meetings, which have been split into designated training, development and evaluation sets. 25 series of meetings have been assigned to the training set, five to the development and five to the test set (table 4). For the purpose of cross-validation, a split into ten parts was also defined ; being this split useful both for ten-fold and five-fold cross-validation.

Subset	Meetings	#meetings	#series
Training set	ES2002, ES2005-2010, ES2012-2016 IS1000-1007 TS3005 TS3008-3012	98	25
Development set	ES2003, ES2011, IS1008, TS3004, TS3006	20	5
Evaluation set	ES2004, ES2014, IS1009, TS3003, TS3007	20	5
All scenario data		138	35

Table 4: The split of the AMI scenario data into training, development and evaluation sets.

3 Previous Work on Automatic Dialogue Act Recognition

The DA recognition task comprises two related sub-tasks: segmentation, and classification or tagging. These tasks may be performed jointly or sequentially. In a sequential approach the conversation is first segmented into unlabelled DA segments, then each detected segment is tagged with a DA label. The joint approach performs both tasks concurrently, detecting DA segment boundaries and assigning labels in a single step. The joint approach is able to examine multiple segmentation and classification hypotheses in parallel, whereas only the most likely segmentation is supplied to the DA classifier in a sequential approach. The joint approach is potentially capable of greater accuracy, since it is able to explore a wider search space, but the optimization problem can be more challenging. In a sequential system the two sub-tasks can be optimised independently. Note that an integrated system may be used as a segmenter by ignoring its classifications. For purposes of comparison, often it may also be used as a classifier, by forcing a human DA segmentation onto it.

Most previous work concerned with DA modelling has focused on tagging presegmented DAs, rather than the overall recognition task which includes segmentation and tagging. Indeed, automatic linguistic segmentation [Stolcke and Shriberg, 1996, Shriberg et al., 2000, Baron et al., 2002] is often regarded as a research problem itself.

3.1 Automatic Dialogue Act Tagging

The use of a generative HMM discourse model [Nagata and Morimoto, 1993], in which observable feature streams are generated by hidden state DA sequences, has underpinned most approaches to DA modelling, and a good overview of this approach is given by Stolcke et al. [2000]. The discourse history is typically modelled using an n-gram over DAs, although approaches such as polygrams [Warnke et al., 1997] have been tested. Lexical features have been widely used for DA tagging (section 3.3), via cue words or statistical language models, including approaches such as multiple parallel n-grams [Venkataraman et al., 2005], hidden event language models [Zimmermann et al., 2006a], and factored language models [Ji and Bilmes, 2005]. Several authors have previously investigated the use of prosody to disambiguate between different DAs with a similar lexical realisation [Bhagat et al., 2003], and investigated approaches to automatically select the most informative features [Shriberg et al., 1998, Hastie et al., 2002]. Prosodic features such as duration, pitch, energy, rate of speech and pauses have been successfully integrated into the processing framework.

Ji and Bilmes [2005] have proposed a switching-DBN based implementation of the HMM approach above outlined, applying it to the DA tagging task on ICSI meeting data. They also investigated a conditional model, in which the words of the current sentence generate the current dialog act (instead of having dialogue acts which generate sequence of words). DA tagging experiments have been performed both using multiple parallel n-grams or adopting a FLM with two factors: words and DA labels. The generative approach prevails over the conditional model, reporting the best classification accuracy when used in conjunction with a FLM. Since this work used only lexical features, and a large number of DA categories (62), a direct comparison with the results provided by [Ang et al., 2005,

Zimmermann et al., 2006a, Dielmann and Renals, 2007a, Zimmermann et al., 2006b] is not possible.

Venkataraman et al. [2003] proposed an approach to bootstrap a HMM-based dialogue act tagger from a small amount of labeled data followed by an iterative retraining on unlabeled data. This procedure enables a tagger to be trained on an annotated corpus, then adapted using similar, but unlabeled, data. The proposed tagger makes use of the standard HMM framework, together with dialogue act specific language models (3-grams) and a decision tree based prosodic model. The authors also advance the idea of a completely unsupervised DA tagger in which DA classes are directly inferred from data.

More recently, there have been a number of conditional models applied to DA classification including support vector machines (SVMs) [Fernandez and Picard, 2002, Liu, 2006] and maximum entropy classifiers [Venkataraman et al., 2005, Ang et al., 2005]. Features for these models include both lexical and prosodic cues, as well as contextual DA information [Venkataraman et al., 2005] (table 5).

A framework for the automatic DA classification of the Spanish CallHome spontaneous speech corpus (using 8 DA labels) has been outlined by Fernandez and Picard [2002]. The proposed approach relies on a SVM based classifier and a set of features derived from energy and pitch contours. Numerical results show the importance of prosodic cues, highlighting how even without a lexical transcription it is still possible to detect DAs well above chance.

Liu [2006] proposed an automatic DA classifier based on the combination of multiple binary SVM classifiers via Error Correction Output Codes. This work extends the 5 DA NIST tagging task outlined in Ang et al. [2005] comparing the originally adopted maximum entropy classifier with a multiclass SVM and 4 different setups based on ECOC SVM classifiers. All ECOC classifiers perform better than a multiclass SVM, but unfortunately they are not able to outperform the baseline MaxEnt system of Ang et al. [2005].

Generative and conditional approaches can also be combined: for example Surendran and Levow [2006] integrated local discriminative SVM classifiers (using prosodic and lexical features) within an HMM framework by applying Viterbi decoding to class posterior probabilities estimated using the SVMs. The SVM-HMM system has been applied to the 13 DA classes Maptask corpus [Carletta et al., 1997] consisting in dialogues between two participants interacting on a game-move task: a *giver* provides instructions to guide a *follower* through the route on a map.

3.2 Automatic Dialogue Act Recognition

An early system for the integrated joint DA segmentation and classification has been outlined by Warnke et al. [1997]. 18 DA classes are automatically recognised in short task oriented two person conversations (appointment scheduling of the German VERBMOBIL corpus). The system using: a multi-layer perceptron and a Language Model for segmentation, a polygram LM for DA classification, and a joint search algorithm to score multiple joint recognition hypotheses; reports an improvement over a sequential approach.

[Ang et al., 2005] addressed the automatic dialogue act recognition problem using a sequential approach, in which DA segmentation was followed by classification of the candi-

date segments. Promising results were achieved by integrating a boundary detector based on *vocal pauses* (table 6) with a hidden-event language model HE-LM (a language model including dialogue act boundaries as pseudo-words). The dialogue act classification task was carried out using a maximum entropy classifier, together with a relevant set of textual and prosodic features. This system segmented and tagged DAs in the ICSI Meeting Corpus (using the 5 broad DA categories outlined in section 2.1), with relatively good levels of accuracy. However results comparing manual with automatic ASR transcriptions indicated that the ASR error rate resulted in a substantial reduction in accuracy.

In a later work Zimmermann et al. [2006a] compared two joint approaches on the same experimental setup. An extended HE-LM able to predict not only DA boundaries but also the type of the DA, and a HMM recogniser inspired by HMM based part of speech taggers, was trained on lexical features and compared using several of the metrics discussed in section 4. The joint HE-LM system obtained lower recognition error rates than the HMM based DA recogniser, achieving performances closer to the discriminative sequential approach of Ang et al. [2005].

A further extension of the joint HE-LM DA recogniser introduced by Zimmermann et al. [2006a] has been developed in Zimmermann et al. [2006b]. A discriminative maximum entropy DA boundary detector and tagger is trained on discretised inter-word pauses with a lexical context of 4 words. Then the weighed combination of the classification probabilities for both systems (HE-LM and MaxEnt) provides the most likely sequence of labelled DA units. Experimental results on the ICSI 5 DA tasks suggest that the novel combined approach is capable of better recognition performances than the sequential approach of Ang et al. [2005]. Note that multiple concurrent DA segmentation and classification hypotheses could be evaluated by joint DA recognisers, enabling the investigation of larger search spaces compared with two-step sequential segmentation-classification approaches.

An integrated framework for the joint DA segmentation and tagging has been outlined by Dielmann and Renals [2007a]. The proposed system is based on: a switching dynamic Bayesian network (DBN) architecture, a set of features related to lexical content and prosody, and a Factored Language Model. The switching DBN coordinates the recognition process by integrating all the available resources. Experiments on the 5 broad DA categories of the ICSI meeting corpus have been carried out, using both manually transcribed speech, and the output of an automatic speech recogniser, and using different feature configurations. The DA segmentation and recognition results are similar to those of Ang et al., although using a discriminative MaxEnt DA classifier [Ang et al., 2005] resulted in a 5% lower error rate for the tagging task. Experiments on the AMI corpus using an extended version of the switching DBN framework have been reported in Dielmann and Renals [2007b].

3.3 Features for Automatic Dialogue Act Processing

Table 5 lists some of the features used in previous works to perform automatic DA classification; while table 6 shows the most frequently used features that have been adopted for the DA segmentation task.

The most common features used for the automatic DA segmentation and classification can be subdivided in:

Feature / Article	Ang et al. [2005]	Rosset and Lamel [2004]	Fernandez and Picard [2002]	Rotaru [2002]	Lendvai et al. [2003]	Andernach [1996]	Reithinger and Klesen [1997]	Venkataraman et al. [2002]	Venkataraman et al. [2003]	Keizer and Akker [2005]	Venkataraman et al. [2005]	Jurafsky et al. [1998]	Zimmermann et al. [2006a]	Zimmermann et al. [2005]	Warnke et al. [1997]	Katrenko [2004]	Webb et al. [2005]	Ji and Bilmes [2005]	Surendran and Levow [2006]	Liu [2006]	Dielmann and Renals [2007b]	Verbree et al. [2006]
Sentence length	✓									✓	✓									✓	✓	✓
First two words	✓	✓									✓									✓		
Last two words	✓										✓									✓		
Number of utterances		✓																				
Bigrams of words in segment				✓																		
Bigram of first two words																				✓		
Utterance type						✓																
Presence/absence Wh-words						✓																
Subject Type						✓																
Specific cue words/phrases						✓						✓				✓						✓
First verb type						✓																
Second verb type						✓																
Question mark						✓																✓
Sparse bag of ngrams																			✓			
Specific patterns										✓												
Grammar pattern										✓		✓										
Polygrams of words							✓								✓							
Factored Language Model																		✓			✓	
Part Of Speech ngrams																						✓
Ngrams of words								✓	✓		✓	✓	✓	✓			✓	✓		✓		✓
First word of next segment	✓										✓									✓		
Speaker (turn) change		✓							✓		✓								✓	✓		
Words in last 10 DA's					✓							✓			✓							
Pitch			✓		✓				✓			✓			✓				✓		✓	
Energy			✓		✓							✓							✓		✓	
Duration			✓		✓				✓			✓			✓				✓		✓	
Pauses					✓				✓			✓			✓						✓	
Rate of speech					✓														✓			
Ngrams of previous DA's								✓	✓		✓			✓				✓	✓		✓	✓
Previous DA hyp. / posteriors		✓								✓												
Next DA										✓												
Previous 10 DAs (from ref.)					✓																	

Table 5: Features used for automatic DA-classification in different studies

Feature / Article	Kolar et al. [2006]	Stolcke and Shriberg [1996]	Lendvai and Geertzen [2007]	Zimmermann et al. [2006b]	Dielmann and Renals [2007b]	Ang et al. [2005]	Zimmermann et al. [2006a]
Segmentation only	✓	✓					
Surrounding Words				✓			
Ngrams of words	✓	✓				✓	✓
Part Of Speech ngrams		✓					
Tokenized Words			✓				
Bag of Words			✓				
Word relevance					✓		
Factored Language Model					✓		
Disfluencies			✓				
Repeats	✓						
Overlapping Speech			✓				
Pauses	✓		✓	✓	✓	✓	
Pitch	✓				✓		
Duration	✓				✓		
Energy	✓				✓		

Table 6: Features used for automatic DA-segmentation in different studies.

Lexical features usually a language model based on words: DA specific ngrams of words, polygrams, factored language models, part-of-speech ngrams, etc. Some systems also rely on selected cue words/phrases and specific lexical or grammatical patterns. The number of words contained by the current DA segment (sentence length) is also a lexical related feature frequently adopted for DA classification. In order to evaluate fully automatic DA tagging and recognition systems, automatic ASR transcriptions are required. Inaccuracies of the automatically recognised speech have an adverse effect on lexical derived features. Therefore it is worth evaluating the full system both on manual and automatic transcriptions in order to estimate the overall degradation of performances caused by the ASR output.

Context features describe the relation between the current and the surrounding utterances, e.g. to indicate temporal overlap between speakers.

Prosodic features represent a wide group of acoustic related features like: F0 and pitch slopes, the duration of words, unvoiced pauses, speech rate, features derived from spectral coefficients, etc.

A discourse model (or discourse grammar) is based on the DA types of the preceding or surrounding segments. It is important to note whether this history is maintained on

the actual output of the DA classifier, or on the hand-annotated DAs. For a realistic evaluation, the actual classification results should be used; however, generating the history from annotated DAs gives an estimation of the potential usefulness of this kind of features.

Two important aspects related to the feature extraction process are source and scope of the extracted features. Even if all the information required for feature extraction should come from fully automatic approaches, several systems are trained on features relying on manually labelled data. Moreover many systems are frequently evaluated using features based on manual annotations (i.e: lexical features estimated using the reference orthographic transcriptions), either because data from an automatic system are not available yet, or to assess the potential usefulness of a new feature family. Automatic DA processing is often a component block of a larger infrastructure (section 5), therefore specific constraints imposed by the applicative domain have a deep influence on the feature scope. For example, in a meeting browsing application designed to offer its facilities online during an undergoing meeting, the DA recognition process will have access only to the past conversations. Note also that in this application the DA processing should operate in real-time relying on a less accurate ASR transcription. In a post-processing application (e.g., offline meeting corpus browser), the whole discourse is available, allowing the use of features which look ahead in the time.

4 Metrics and Evaluation

Each of the segmentation, classification and the joint segmentation and classification tasks, has its own set of performance metrics. If performance evaluation is straightforward for the DA tagging task, the same cannot be said about DA segmentation or recognition tasks. Several evaluation metrics have been proposed, but the debate on this topic is still open. Moving from the NIST-SU error metric introduced in NIST website [2003], several DA segmentation and recognition metrics have been proposed by Ang et al. [2005] and subsequently extended by Zimmermann et al. [2006a].

4.1 Classification metrics

The performance of DA classification using manually annotated segments is usually measured in terms of accuracy, which is the percentage of correctly classified segments, or classification error rate, which is the percentage of incorrect classifications. For a more detailed evaluation, occurrences and correct classifications of each DA class can be counted separately [Lesch et al., 2005a]:

$$
\begin{aligned}
correct_{DA} &= \textit{the number of times DA was correctly classified} \\
annotated_{DA} &= \textit{the number of occurrences of DA in the annotated test data} \\
tagged_{DA} &= \textit{the number of times DA was classified}
\end{aligned}
$$

Based on these counts, we define the recall ($Recall_{DA}$) and precision ($Precision_{DA}$) measures for each DA class, as well as the accuracy and mean precision for the whole test

set:

$$
\begin{aligned}
Recall_{DA} &= \frac{correct_{DA}}{annotated_{DA}} \\
Precision_{DA} &= \frac{correct_{DA}}{tagged_{DA}} \\
Accuracy &= \frac{\sum_{DA} correct_{DA}}{\sum_{DA} annotated_{DA}} \\
Precision &= \frac{\sum_{DA} Precision_{DA} * annotated_{DA}}{\sum_{DA} annotated_{DA}}
\end{aligned}
$$

4.2 Segmentation metrics

The evaluation of the automatic DA segmentation is a non-trivial task. Several evaluation metrics can be defined, each giving a different perspective on the segmentation results. Figure 1 illustrates the principal metrics used to evaluate the accuracy of automatic DA segmentation. NIST-SU, recall, precision, f-measure and boundary are based on boundaries. Each word is followed by a potential boundary position, and segmentation is a binary classification into boundaries and non-boundaries. There are four possible outcomes: boundaries may be correctly identified (true positives, tp) or missed (false negatives, fn), non-boundary positions may be correctly identified (true negatives, tn) or a false boundary may be hypothesised (false positives, fp). The sum $tp + tn + fp + fn$ is equal to the number of words. The occurrences of these four events are counted. The boundary-based metrics take different combinations of these counts into consideration:

$$
\begin{aligned}
NIST - SU &= \frac{fp + fn}{tp + fn} \\
Boundary &= \frac{fp + fn}{tp + tn + fp + fn} \\
Recall &= \frac{tp}{tp + fn} \\
Precision &= \frac{tp}{tp + fp}
\end{aligned}
$$

The F-measure is the harmonic mean of the computed precision and recall given the reference sentence boundaries and the boundaries hypothesised by the segmentation system: $F = 2 \times Recall \times Precision/(Recall + Precision)$. The other two segmentation metrics, DA segment error rate (DSER) and Strict, are based on segments. DSER is the fraction of reference segments which have not been correctly recognised, meaning that either of the boundaries is incorrect. Strict is a variant of DSER in which each DA segment is weighted with its length (number of words).

4.3 Joint segmentation and classification metrics

The DA recognition task is more challenging, since the limited accuracy of automatic segmentation and classification are combined together. Note that a direct comparison between DA recognition and classification results is difficult. However the DA classification

Reference	S|Q.Q.Q.Q|S.S.S|B|S.S|
System	S|Q|S|Q.Q|D.D.D|S.S|S|
NIST-SU	.c.e.e...c.....c.e.e.c
Boundary	.c.e.e.c.c.c.c.c.e.e.c
Recall	.c.......c.....c.e...c
Precision	.c.e.e...c.....c...e.c
DSER	c|...e...|..c..|e|.e.|
Strict	c|e.e.e.e|c.c.c|e|e.e|

Metric	Counts	Reference	Rate
NIST-SU	3 FP, 1 miss	5 boundaries	80%
Boundary	3 FP, 1 miss	11 (non-)boundaries	27%
Recall	4 correct	5 boundaries	80%
Precision	4 correct	7 hypothesised boundaries	57%
F-Measure	-	-	67%
DSER	3 match errors	5 reference DAs	60%
Strict	7 match errors	11 reference words	63%

Figure 1: Metrics for segmentation based on boundaries (NIST-SU, Recall, Precision, F-Measure and Boundary) and on segments (DSER and Strict). The symbol '|' is used to indicate boundaries between consecutive DAs and '.' stands for non-boundaries between words. The letters S, Q, D, and B represent single words of the DAs. Correctly hypothesised boundaries are marked with a letter c while e is used to label false positives and missed boundaries.

performance can be interpreted as an upper boundary for the whole recognition process, which would be reached if automatic segmentation was perfect.

A set of metrics, in analogy to the segmentation metrics of section 4.2, can be defined for the recognition task. Figure 2 illustrates a set of performance metrics for joint segmentation and classification of DAs. In contrast to the NIST error metric for segmentation, the hypothesised DA label is taken into account as well, leading not only to false positives (insertions) and misses (deletions) but also to substitutions. While the strict error metric requires correct DA boundaries the lenient metric completely ignores segmentation errors. As the DER can also be defined via a DA based recall, DA based precision can be defined as well, leading to a DA based F-measure: $F = 2 \times Recall \times Precision / (Recall + Precision)$. Note that recall, precision and F-measure are based on dialogue act units, not on DA boundaries as it was for the segmentation metrics.

While higher values for Recall, Precision and the F-measure indicate higher performances, the remaining metrics are error metrics, thus higher values imply lower performances. It is important to note that these metrics and all evaluations presented in this chapter are intrinsic, being purely based on the comparison between human annotation and classifier/recogniser output. Knowledge of the discourse structure could be beneficial in several applicative domains (section 5); thus the automatically classified/recognised DAs often form the input of further processing stages. However the effects of DA segmentation errors and DA misclassifications on the overall system performances depend on how the DA recogniser output was used. These effects are not taken into account by the metrics defined

Reference	S|Q.Q.Q.Q|S.S.S|B|S.S|
System	S|Q|S|Q.Q|D.D.D|S.S|S|
NIST	.c.e.e...c.....e.e.e.c
Strict	c.e.e.e.e.e.e.e.e.e.e.
Lenient	c.c.e.c.c.e.e.e.e.c.c.
DER/Recall	c|...e...|..e..|e|.e.|
Precision	c|e|e|.e.|..e..|.e.|e|

Metric	Counts	Reference	Rate
NIST	3 FP, 1 miss, 1 subst.	5 boundaries	100%
Strict	10 words	11 words	91%
Lenient	5 words	11 words	45%
DER	4 erroneous dialog acts	5 dialog acts	80%
Recall	1 correct dialog act	5 dialog acts	20%
Precision	1 correct dialog act	7 dialog acts	14%
F-Measure	-	-	17%

Figure 2: Metrics for joint segmentation and classification: the boundary based NIST error rate, the word based strict and lenient metrics, as well as the DA error rate (DER). The DA based recall, precision, and corresponding F-measure are illustrated in the lower part of table. The symbol '|' is used to indicate boundaries between consecutive DAs and '.' stands for non-boundaries between words. The letters S, Q, D, and B represent single words of the same DA unit; S, Q, D, and B also represent the dictionary of 4 possible DA labels. Correctly recognised elements are marked with a letter c while e is used to mark errors.

in table 1 and 2, and are not examined here. Ideally, the users of a DA segmenter/classifier should separately investigate the effects of different DA recognition errors. Given such analysis, the most appropriate metric can be identified, and the DA recognition system can be optimised for this specific application.

4.4 Evaluation on Automatic Speech Recogniser output

The reference DA annotation is produced on top of the manually transcribed word sequence. When the reference orthographic transcription is replaced by the ASR output, the DA tags need to be applied to a different word sequence, owing to ASR errors. Since a manual re-annotation of the ASR output would be extremely expensive, the evaluation scheme proposed by Ang et al. [2005] is often adopted: ASR words are mapped into the manually annotated segments according to their midpoint $0.5 * (word\_start\_time + word\_end\_time)$, thus inheriting their reference DA labels.

Insertions and deletions Since the proposed alignment method is segment-based, insertions and deletions of single words are ignored. However, insertions and deletions of entire DA segments occur if the recogniser finds words outside of the boundaries of any annotated dialogue act, or if no words are recognised within the boundaries of an annotated DA. For example in the AMI meeting corpus, automatic transcriptions are available for 101585 dialogue acts; the *midpoint alignment* results in 91537 annotated dialogue

act segments with recognised words, and 9968 empty DA segments without words. Although this is a large fraction, the information loss is likely to be less severe, as 66% of the deleted segments contain only laughs, coughs and other non-speech noises; 70% are of type Fragment and have no function in the discourse. While 49.2% of the segments of type Fragment are deleted, the loss on all other types is less severe, between 1% and 7%. Only 14% of the deleted segments are non-Fragments containing more than one word.

The deleted DA segments can be considered in three different ways:

Include deletions as misclassifications Often in the ASR output there is no indication that a dialogue act has taken place unless words from it were recognised. Therefore deleted segments will be scored as errors.

Classify deletions However through automatic Speaker Activity Detection it is possible to estimate if a participant spoke, even when no words were recognised by the ASR system. Therefore it is possible to include these segments as ordinary dialogue acts without words. They can be classified using non-lexical features like the duration, overlap with previous DAs, or prosody related features. Classifiers which are limited to lexical features can choose the most frequent class (or the most frequently lost class, e.g. Fragment). Note that this type of evaluation allows a closer comparison to results on manual transcriptions.

Exclude deletions Deletions can be excluded from the accuracy metrics, showing the potential of the DA classifier on ASR words more clearly.

Impact of ASR on DA classification DA tagging experiments both on ICSI [Ang et al., 2005, Dielmann and Renals, 2007a] and AMI data show that the classification accuracy on automatically recognised words is approximately 7–10% (absolute) lower than on reference transcriptions.

5 Applications of Automatic Dialogue Act Processing

Dialogue acts form a useful level of representation for the interpretation of conversations, providing a bridge between an orthographic (word-level) transcription, and a richer representation of the discourse. DA labels may incorporate syntactic, semantic and pragmatic factors: in addition to providing information about the structure of a dialogue and the course of a conversation, DAs are also able to capture, at a coarse level, individual speaker attitudes and intentions, their interaction role and their level of involvement. The reliable recognition of the DA sequence in a conversation, and the resulting knowledge of the discourse structure, can be beneficial in the development of applications in a multitude of domains, such as: spoken dialogue systems, machine translation, automatic speech recognition, automatic summarisation, topic segmentation and labelling, action items detection, group action detection, participant influence detection, and dialogue structure annotation.

As outlined in section 2, during the last decade, multiple corpora have been annotated in terms of DAs, and a relevant literature about automatic DA recognition (section 3) has been developed. Several works also focused on the exploitation of the automatically extracted DAs. Moving from the idea that the knowledge about the ongoing conversation

(conveyed by DAs) can be used to enhance language modelling; improving Automatic Speech Recognition of conversational speech was one of the first targets. Jurafsky et al. [1997a] investigated the use of automatically detected Dialogue Acts to improve Automatic Speech Recognition. The 1155 pre-segmented conversations from the Switchboard database were automatically tagged using the clustered dictionary of 42 DA labels. The system made use of a generative DA tagging infrastructure based on: prosodic features (pitch, speaking rate, energy, etc.), 42 word sequence based trigram models, and a bigram discourse language model. Automatic transcriptions were generated through ASR and then fed to the automatic DA tagger. The automatically detected DA classes are then used to rescore the ASR output by means of a novel *DA conditioned mixture Language Model*: N-best lists associated to each test-set utterance have been rescored using a mixture of DA specific LMs. Numerical results on the Switchboard corpus show only a limited improvement (0.3%) on the ASR word error rate because of the skewed distribution of DA classes (statements account for 83% of the corpus). However DA rescored ASR should have a larger impact on specific tasks with more even DA distributions (e.g., task oriented dialogs). A deeper analysis and further generalisations (*mixture of posteriors*) of the *mixture of language models* have been reported in Stolcke et al. [2000]. Related experiments on Maptask [Taylor et al., 1998], show that the automatic choice of the most appropriate language models from a set of 12 DA specific LMs (detected using intonation modeling), can improve the speech recognition word error rate by an absolute 1%.

Machine translation is another applicative domain where DA recognition can be invaluable, since DAs can help resolve ambiguities in translating utterances. The VerbMobil project investigated machine translation in dialogue systems [Küssner, 1997, Wahlster, 2000], similarly to the work independently done by Lee et al. [1997]. The use of DAs for machine translation of spoken task-oriented dialogues has been also proposed in the context of the C-STAR project by Levin et al. [2003].

Automatic detection of *action items*, intended as public commitments to perform a defined task, is a novel research topic which share some analogies with and relies on automatic DA recognition. In the work of Purver et al. [2007], 4 task specific Action Item Dialogue Acts (description, time-frame, owner and agreement) are automatically detected combining 4 independent SVM classifiers trained on: lexical, prosodic features and conventional ICSI DA tags. The automatically detected AIDAs are then rule-based parsed and summarised in order to outline the identified action items. Disambiguating the pronoun *you*, between its generic and referential use in a conversation, is a task related to *action items* detection, which could be useful to identify the owner of an action item (who committed to perform a given task). The SVM based system proposed by Gupta et al. [2007b], based on DAs, lexical and part of speech features, is able to disambiguate the two uses with an accuracy of 84.4% on 2 person conversations from the Switchboard corpus. This represents a significant result, well above the baseline 56.4% achievable always predicting the dominant class. In particular DAs proved to be crucial for this task, reaching an accuracy of 80.92% even if used alone. Later experiments [Gupta et al., 2007a], using a similar setup on the AMI corpus, reported an accuracy of 75.1% with the full feature setup and 71.9% using only DAs (dominant class baseline of 57.9%).

Automatically detecting when decisions are reached during a conversation is another target application for automatic DA extraction. Hsueh and Moore [2007] used both DA unit

temporal boundaries and DA labels for automatic decision detection in conversational speech. The manually annotated DA units are classified as decision making DAs or non-decision DAs using a MaxEnt classifier and a rich set of lexical, prosodic, topical and contextual features (like speaker role and DA labels). Experiments on the AMI corpus show that decision making DAs can be detected with a precision of about 72% (66% using only contextual features).

Differently from written text, automatically transcribed speech lacks of a proper punctuation. It is often unpractical to process the entire raw transcription or to evaluate the resulting system on unsegmented data, thus shorter speech segments need to be defined. The temporal boundaries of automatically recognised DA units provide a principled way to segment conversational speech. For example Murray et al. [2006] and Murray and Renals [2006] adopted the DA segments as the atomic unit for automatic extractive summarisation; features like lexical cues, speaker activities and term frequencies were individually extracted from each DA unit, and Singular Value Decomposition carried out on the resulting DA based feature vectors. Note that although DA segments are a good solution for automatic speech segmentation, some low-level segmentation techniques such as "Spurts" [Baron et al., 2002], continuous speech segments separated by at least half a second of silence, could represent a viable option.

Complex integrated applications based on automatic DA processing are being currently investigated. For example, topic segmentation and extractive summarisation have been combined in the "AMI Meeting Facilitator" system [Murray et al., 2007], a visual application focused on supporting offline meeting browsing. Here dialogue acts, being exploited by both subtasks (segmentation and summarisation), offer a common ground for the whole system.

6 DA Tagging, Segmentation and Recognition of the AMI Meeting Corpus

The DA tagging and recognition experiments conducted on the AMI meeting corpus extend and adapt the previous experiences acquired on former multiparty conversational corpora, like the ICSI meeting corpus. In order to compare DA classification performances on different meeting data (Switchboard, ICSI and AMI) a portable DA tagger has been developed by Verbree et al. [2006]. The proposed system makes use of several feature families: question marks and lexical cues, unit lengths, compressed ngrams of both words and POS tags; and a bigram discourse model. The extracted features are then modelled using the J48 classifier of the Weka toolkit. While the classification accuracy achieved on the Switchboard 42 DA task is about 5% lower than the state of the art, the system outperforms all the previous works on the 5 DA ICSI task, reaching an accuracy of 89.3%. The classification accuracy on the AMI 15 DA is about 59.8% using reference orthographic transcriptions and 49.3% using the ASR output.

The maximum entropy (MaxEnt) based classification system outlined in [Lesch, 2005, Lesch et al., 2005b] adopts a wide set of features belonging to the following 5 classes: lexical features, DA unit length and duration, temporal relation between adjacent utterances, speaker change and dialogue act history. A feature selection algorithm, which grows the

	Metric	Reference	ASR output
S	NIST-SU	20.4	26.5
E	DSER	12.8	17.0
G	Strict	28.5	29.4
M.	Boundary	3.1	4.4
R	NIST-SU	71.3	85.9
E	DER	51.9	62.5
C.	Strict	62.1	68.5
	Lenient	42.2	48.3

Table 7: DA segmentation and recognition error rates (%) on the AMI meeting corpus both on reference manual transcriptions and ASR output; segmentation results are reported using the interpolated FLM, whenever the hybrid FLM+iFLM system has been used for the joint DA recognition task.

feature subset by iteratively ranking the features according to their classification accuracy, has been adopted to select only the most relevant features and reduce the feature set. The best classification accuracy obtained on the AMI evaluation set is 65.8% for reference words and 54.9% with automatically recognised words (classifying ASR deleted DA units by chance). This result defines the state of the art for the 15 DA AMI tagging task. Similarly to Verbree et al. [2006] and Dielmann and Renals [2007b], when the reference transcription is replaced by the ASR output, the classification accuracy falls by about 10% (absolute).

The discriminative MaxEnt approach outperforms the generative FLM based classifier of Dielmann and Renals [2007b] by about 6% both on reference (59.1%) and automatic transcriptions (49.3%). However the switching DBN infrastructure outlined in Dielmann and Renals [2007b], being able to perform concurrently both DA segmentation and classification, is principally targeted to the joint DA recognition task rather than being forced to classify presegmented data. DA recognition experiments have been reported using three different language model configurations: an FLM trained only on AMI data, a weighted interpolated FLMs trained also on ICSI and Fisher data, and an hybrid setup with both an FLM and an interpolated FLM. The interpolated FLM, thanks to its richer dictionary and language model, reduces the number of segmentation errors by a factor of 2–3, at the cost of a slightly degraded DA classification accuracy. A hybrid approach, using both FLMs, allows a trade off between segmentation and classification, improving the overall recognition accuracy. Note also that joint DA recognition approaches perform segmentation and classification in a single and indivisible process, such that adjustments which improve the segmentation may lead to lower classification accuracy and vice-versa. The reported experiments (table 7) suggest that it is possible to perform automatic segmentation into DA units with a relatively low error rate. However the operation of automatic recognition into 15 imbalanced DA categories has a relatively high error rate, indicating that this remains a challenging task.

Both DA tagging and automatic DA recognition are open research topics, thus further investigations and improvements both on the feature extraction process and on the statistical modelling framework will be discussed in the next paragraphs.

Features Language models automatically derived from text corpora are typically very large, with up to several hundred thousand n-gram features. The system outlined by Verbree et al. [2006] presents an approach to select a small number of lexical cues, which shows relatively good classification accuracies even using very small models. Shrinking the feature set helps reducing the computing overhead when a DA recogniser will be employed as part of an online application, deemed to run in real-time. In many cases a smaller model with 1-2% lower accuracy, which fits easily on a machine together with various other modules, will be preferable to a slightly better model with vast memory requirements.

Likewise, the selection of feature types investigated using a maximum entropy modelling approach [Lesch, 2005, Lesch et al., 2005b], reduces the number of binary features and provides invaluable insights to the usefulness and the overlap between different types of features. Lexical features, utterance length and duration, as well as context-dependent features positively contribute to the final results. Being the lexical features, especially word identities, utterance-initial and utterance-final words, the most salient ones.

Future research should include an even deeper analysis of the individual contributions granted by the current features; and should examine the potential of introducing new feature families:

Multi-modal features Other modalities like gestures or focus of attention may provide additional valuable clues.

Forward-looking features The experiments conducted so far on the AMI meetings make use only of backward-looking features, i.e. features derived from material up to the current utterance. Assuming that the entire meeting is available and the DA recognition framework will be part of an application which allows offline processing, nothing prevents from exploiting forward-looking features such as: word identities from the following utterance, future speaker changes, etc.

Additional information from speech activity detection Some utterances are lost in the ASR recognition process, since none of their words were recognised. However, some of these utterances can probably be recovered using automatic speaker activity detection (section 4.4).

Advanced classification methods The DA classification methods applied to the AMI task are based on "flat" models which discriminate between all 15 DA types in one step. However it is possible to combine multiple specialised classifiers creating a hierarchically layered classifier.

Models with fewer classes are often more discriminative than models with a large number of classes. One way to take advantage of this is to group the classes and perform the final classification in two or more steps. DA types which are similar, or frequently confused, can be merged into one abstract class, resulting in a model with fewer classes. When this model predicts an abstract class, a secondary model can be used, which is trained to discriminate between the subclasses which were collapsed into the abstract class.

Another approach targeted on reducing the number of classes is based on the notion of dialogue act dimensions: each DA type can be described by a tuple of dimensions, each of which has a small set of values. Each dimension can be modeled separately, and a

meta model can be applied to map a tuple of values to an actual DA type. While the dialogue act labels of the ICSI-MRDA scheme are clearly composed of one or more tags which represent orthogonal properties of the utterance, the AMI DA scheme consists only of a flat list of 15 DA types. However, it is still possible to identify various aspects, or "dimensions", in which any two of the 15 types are similar or different. For instance Elicit-Inform, Elicit-Offer-Or-Suggestion, etc. have in common that they elicit information from the other participants. On the other hand, Offer and Elicit-Offer-Or-Suggestion are similar in that both of them are concerned with offers. Thus we can hypothesise that two aspects of the AMI dialogue acts are: "information type" (inform, suggest/offer, assess, ...) and "direction" (whether the speaker expresses information, or elicits information).

References

T. Andernach. A machine learning approach to the classification of dialogue utterances. *Computing Research Repository*, July, 1996.

A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. The HCRC map task corpus. *Language and Speech*, 34:351–366, 1991.

J. Ang, Y. Liu, and E. Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. In *Proc. ICASSP*, volume 1, pages 1061–1064, Philadelphia, USA, 2005.

J. L. Austin. *How to do Things with Words.* Oxford: Clarendon Press, 1962.

E.G. Bard, C. Sotillo, A.H. Anderson, H.S. Thompson, and M.M. Taylor. The DCIEM map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment. *Speech Communication*, 20(1):71–84, 1996.

D. Baron, E. Shriberg, and A. Stolcke. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *ICSLP*, Denver, Colorado, USA, September 2002.

S. Bhagat, H. Carvey, and E. Shriberg. Automatically generated prosodic cues to lexically ambiguous dialog acts in multiparty meetings. In *Proc. International Congress of Phonetic Sciences*, pages 2961–2964, August 2003.

J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, In Press.

J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. The reliability of a dialog structure coding scheme. *Computational Linguistics*, 23: 13–31, March 1997.

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 meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*, 2005. AMI-108.

A. Dielmann and S. Renals. Multistream recognition of dialogue acts in meetings. In *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–06)*, pages 178–189. Springer, 2007a.

A. Dielmann and S. Renals. DBN based joint dialogue act recognition of multiparty meetings. In *Proc. IEEE ICASSP*, volume 4, pages 133–136, April 2007b.

R. Fernandez and R.W. Picard. Dialog act classification from prosodic features using support vector machines. In *Proceedings of speech prosody 2002*, April 2002.

J Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520, San Francisco, March 1992.

S. Gupta, J. Niekrasz, M. Purver, and D. Jurafsky. Resolving "you" in multi-party dialog. In *SIGdial*, September 2007a.

S. Gupta, M. Purver, and D. Jurafsky. Disambiguating between generic and referential "you" in dialog. In *ACL*, June 2007b.

H. Hastie, M. Poesio, and S. Isard. Automatically predicting dialogue structure using prosodic features. *Speech Communication*, (36):63–79, 2002.

Y. Horiuchi, Y. Nakano, H. Koiso, M. Ishizaki, H. Suzuki, M. Okada, M. Naka, S. Tutiya, and A. Ichikawa. The design and statistical characterization of the japanese map task dialogue corpus. *Journal of Japanese Society for Artificial Intelligence*, 14(2):261–272, 1999.

P. Hsueh and J. Moore. What decisions have you made: Automatic decision detection in conversational speech. In *NACCL/HLT*, pages 25–32, Rochester, NY, USA, April 2007.

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, Hong Kong, China*, pages 364–367, April 2003.

G. Ji and J. Bilmes. Dialog act tagging using graphical models. In *Proc. ICASSP*, volume 1, pages 33–36, Philadelphia, USA, 2005.

D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Ess-Dykema. Automatic detection of discourse structure for speech recognition and understanding. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 88–95, Santa Barbara, CA, US, 1997a. IEEE CS.

D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation (coders manual, draft 13). Technical report, Univ. of Colorado, Inst. of Cognitive Science, 1997b. URL `http://www.icsi.berkeley.edu/cgi-bin/pubs/publication.pl?ID=001359`.

D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. Lexical, prosodic, and syntactic cues for dialog acts. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pages 114–120. Association for Computational Linguistics, Somerset, New Jersey, 1998. URL `citeseer.ist.psu.edu/article/jurafsky98lexical.html`.

S. Katrenko. Textual data categorization: back to the phrase-based representation. In *Proceedings in 2nd International IEEE Conference "Intelligent systems", Vol. III*, pages 64–67, June 2004.

S. Keizer and R. op den Akker. Dialogue act recognition under uncertainty using bayesian networks. *Natural Language Engineering*, 1:1–30, 2005.

M. Klein, N. Ole Bernsen, S. Davies, L. Dybkjær, J. Garrido, H. Kasch, A. Mengel, V. Pirrelli, M. Poesio, S. Quazza, and C. Soria. Supported coding schemes. Technical Report MATE Deliverable D1.1, EU project LE4-8370, 1998. URL `http://mate.nis.sdu.dk/about/D1.1/`.

J. Kolar, E. Shriberg, and Y. Liu. Using prosody for automatic sentence segmentation of multi-party meetings. In *Proc. TSD 2006*, volume 9, pages 629–636, 2006.

U. Küssner. Applying dl in automatic dialogue interpreting. In *International Workshop on Description Logics*, pages 54–58, Yvette, France, 1997.

J. Lee, G. C. Kim, and J. Seo. A dialogue analysis model with statistical speech act processing for dialogue machine translation. In *Spoken Language Translations EACL97 Workshop*, pages 10–15, Budapest, Hungary, 1997.

P. Lendvai and J. Geertzen. Token-based chunking of turn-internal dialogue act sequences. In *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue*, pages 174–181, Antwerp, Belgium, 2007.

P. Lendvai, A. van den Bosch, and E. Krahmer. Machine learning for shallow interpreta-

tion of user utterances in spoken dialogue systems. In *Proceedings of EACL-03 Workshop on Dialogue Systems: interaction, adaptation and styles of management*, pages 69–78, 2003.

S. Lesch. Classification of Multidimensional Dialogue Acts using Maximum Entropy. Diploma thesis, Saarland University, Postfach 151150, D-66041 Saarbrücken, Germany, December 2005.

S. Lesch, T. Kleinbauer, and J. Alexandersson. A new Metric for the Evaluation of Dialog Act Classification. In *Proceedings of the Ninth Workshop On The Semantics And Pragmatics Of Dialogue (SEMDIAL 2005) – DIALOR'05*, pages 143–146, Nancy, France, June 2005a.

S. Lesch, T. Kleinbauer, and J. Alexandersson. Towards a Decent Recognition Rate for the Automatic Classification of a Multidimensional Dialogue Act Tagset. In *Proceedings of the 4^{th} IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 46–53, Edinburgh, Scotland, UK, August 2005b.

L. Levin, C. Langley, A. Lavie, D. Gates, D. Wallace, and K. Peterson. Domain specific speech acts for spoken language translation. In *SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, 2003.

Y. Liu. Using SVM and error-correcting codes for multiclass dialog act classification in meeting corpus. In *Proc. Interspeech - ICSLP*, pages 1938–1941, September 2006.

G. Murray and S. Renals. Dialogue act compression via pitch contour preservation. In *Interspeech*, Pittsburgh, USA, September 2006.

G. Murray, S. Renals, J. Carletta, and J. Moore. Incorporating speaker and discourse features into speech summarization. In *NACCL/HLT*, pages 367–374, New York, USA, June 2006.

G. Murray, P. Hsueh, S. Tucker, J. Kilgour, J. Carletta, J. Moore, and S. Renals. Automatic segmentation and summarization of meeting speech. In *NACCL/HLT*, pages 9–10, Rochester, NY, USA, April 2007.

M. Nagata and T. Morimoto. An experimental statistical dialogue model to predict the speech act type of the next utterance. *Proc. of the International Symposium on Spoken Dialogue*, pages 83–86, November 1993.

NIST website. Rt-03 fall rich transcription. http://www.nist.gov/speech/tests/rt/rt2003/fall/, 2003.

M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, and S. Noorbaloochi. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, september 2007.

N. Reithinger and M. Klesen. Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*, pages 2235–2238, Rhodes, Greece, 1997.

S. Rosset and L. Lamel. Automatic detection of dialog acts based on multi-level information. In *Procedings of the ICSLP*, pages 540–543, Jeju Island, Korea, October 2004. URL `ftp://tlp.limsi.fr/public/TuB401o.2\_p540.pdf`.

M. Rotaru. Dialog act tagging using memory-based learning. Technical report, University of Pittsburgh, spring 2002. Term project in Dialogue-Systems class.

J. Searle. *Speech Acts*. Cambridge University Press, 1969.

E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, (41):439–487, 1998.

E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32:127–154, September 2000.

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, Cambridge, MA, USA*, pages 97–100, Cambridge, USA, April-May 2004.

A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In *Proc. ICSLP*, volume 2, pages 1005–1008, October 1996.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373, 2000. URL `citeseer.ist.psu.edu/stolcke00dialogue.html`.

D. Surendran and G. A. Levow. Dialog act tagging with support vector machines and hidden Markov models. In *Proc. Interspeech - ICSLP*, September 2006.

P. Taylor, S. King, S. Isard, and H. Wright. Intonation and dialog context as constraints for speech recognition. *Language and Speech*, 41:489–508, 1998.

A. Venkataraman, A. Stolcke, and E. Shirberg. Automatic dialog act labeling with minimal supervision. In *Proceedings of the 9th Australian International Conference on Speech Science & Technology*, December 2002.

A. Venkataraman, L. Ferrer, A. Stolcke, and E. Shriberg. Training a prosody-based dialog act tagger from unlabeled data. *Proc. of the IEEE ICASSP*, April 2003.

A. Venkataraman, Y. Liu, and E. Shriberg. Does active learning help automatic dialog act tagging in meeting data? In *Proc. Interspeech - Eurospeech*, pages 2777–2780, September 2005.

D. Verbree, R. Rienks, and D. Heylen. Dialogue-act tagging using smart feature selection; results on multiple corpora. In *IEEE Spoken Language Technology Workshop*, pages 70–73, December 2006.

W. Wahlster. *Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System*, pages 3–21. Springer, 2000.

V. Warnke, R. Kompe, H. Niemann, and E. Nöth. Integrated dialog act segmentation and classification using prosodic features and language models. In *Proc. 5th Europ. Conf. on Speech, Communication, and Technology*, pages 207–210, September 1997. Eurospeech.

N. Webb, M. Hepple, and Y. Wilks. Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, 2005.

M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke. A* based joint segmentation and classification of dialog acts in multiparty meetings. In *Proc. 9th IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 215–219, San Juan, Puerto Rico, november 2005.

M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke. Toward joint segmentation and classification of dialog acts in multiparty meetings. In *Machine Learning for Multimodal Interaction: 2nd International Workshop, MLMI 2005*, pages 187–193. LNCS 3869, Springer, 2006a.

M. Zimmermann, A. Stolcke, and E. Shriberg. Joint segmentation and classification of dialog acts in multiparty meetings. In *Proc. IEEE ICASSP*, volume 1, May 2006b.

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

Meeting Browsing

November 7, 2007

AMI Consortium State of the Art Report

Meeting Browsing

November 7, 2007

Abstract

A meeting browser is an application designed to allow users to access archived meeting recordings. Though browsing might figure heavily when accessing such archives, the application should also support search and any other interactions that could take place between an end user and a meetings archive. This document examines the state of the art of meeting browsers. We begin by re-classifying browsers and related applications into three tiers according to the source of the data they primarily make use of. We look at each tier and discuss the problems faced at each tier and the solutions designed to address these problems. We then examine two browsers in detail - one which offers a complete recording and browsing system, and a meta browser which allows the user to select which components they want to use. We then conclude by briefly examining the process of evaluating meeting browsers.

1 Introduction

Given the ever decreasing cost of capture and storage of multimedia data the recording and archiving of meetings is now relatively common. To access these archives a *meeting browser* is typically used. Despite its name, any application which acts as a front end to a meeting archive is considered to be a meeting browser whether the primary focus is on browsing, search, summarisation, or other forms of interaction. Meeting browsing is an emerging field but despite this there are a large numbers of browsers described in the literature.

To organise meeting browser research this report refines the classification scheme described in Tucker and Whittaker [2005] and Bouamrane and Luz [2007]. There browsers were separated into groups according to the type of data they made primary use of for navigation or presentation. Four groups were selected: audio, video, artefact and discourse browsers (Although Bouamrane and Luz [2007] analyse the first three groups only). In this document we refine this classification by separating browsers into tiers (see Fig. 1). The first tier comprises data *recorded* during the meeting - namely the audio and video recordings. The second tier consists of data that the *participants create* during the meeting - personal notes, slides, minutes etc. The third tier consists of browsers that make primary use of data *derived* from the previous two tiers - speech transcripts, focus of attention, higher level annotations etc. Note that browsers generally also make use of data from lower tiers.

This report continues by examining the requirements of a meeting browser, then analyses meeting browsers that have been developed in the past using this modified taxonomy.We then examine two browsers in detail and conclude by briefly examining how browsers have been evaluated.

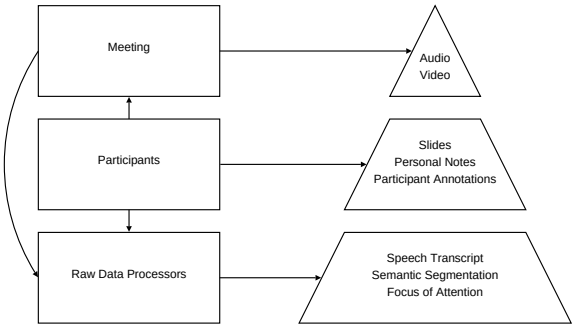


Figure 1: The three tiers of meeting data.

2 Motivations

Little work has examined what users require from meeting browsers but generally one of three methods has been used in order to elicit requirements: Large scale surveys of tools and memory processes, query elicitation and analysis of current practices. These studies are discussed in detail in Whittaker et al. [2008] but we briefly revisit the major findings here.

Jaimes et al. [2004] used a questionnaire to assess the current use of tools to review meetings, finding that meeting participants often reconstructed meeting information rather than remembering it verbatim. A similar approach was used by Lisowska [2003] (and Lisowska et al. [2004a]) who asked participants to generate questions that they would ask of a meeting browser.

Whittaker et al. [1994] examined current recording practices by interviewing people who currently recorded meetings and investigating what the advantages and disadvantages of their recording approaches were. A second interview study investigated the note-taking practices of meeting participants and sought to identify and address the problems that note-taking generates. Additionally, Whittaker et al. [2008] describes an ethnographic study of two firms which investigated the problems of benefits of both personal (notes) and public (minutes) meeting records.

These studies combined have implications for browser designers. Firstly the studies found that users of meeting browsers frequently had the need to access abstractions rather than raw data. Thus, when accessing meeting archives users rarely want to review an entire meeting but instead need to focus on short, relevant parts of the meeting. Secondly users expressed a desire to access meeting archives through categorisations that were relevant to them - e.g. agenda items, decisions and actions. These types of data are often poorly supported by meeting browsers. The overall finding of these studies suggests that browsers are currently too complex and unfocused with regard to user requirements. The two third tier browsers we discuss in detail below go some way towards addressing these criticisms.

3 Meeting Browsers

3.1 First Tier Browsers

Since the first tier browsers are focused on raw multimedia recordings it is easier to examine how the problems of navigating to and locating relevant information have been investigated. Therefore to examine this tier of browsers we look at the solutions to the problems associated with audio rather than at specific applications.

3.1.1 Speech Marking

Degen et al. [1992] exemplifies the *audio marking* approach. Such systems allow users to manually mark relevant points of the meeting as it is recording, here by using a 'marker' button on the recorder itself. When playing back the recording the markers can be used as a means of navigating to points of interest. The recording interface has two different marker buttons which can be distinguished when playing back the recording.

3.1.2 Speech Segmentation

As well as allowing users to manually mark (and therefore manually segment) speech recordings other systems have investigated *automatic* segmentation of speech in order to aid navigation. The means by which the recording is segmented varies depending on the application. For example, Hindus and Schmandt [1992] look at segmenting informal workplace discussions according to both person and also by pauses that each person makes between utterances. Other possibilities for segmentation can be more semantic or can make use of prosodic cues (e.g. Arons [1997]) in order to allow the user to navigate the audio recording more efficiently.

3.1.3 Playback mechanisms

Segmenting the recording goes some way to giving users near-random access to speech recordings. However the problem of processing relevant audio still remains - whilst speech is easy to record we are required to listen at around 150 words per minute, whereas we can read at 600 words per minute (and are adept at skimming text to locate regions of interest). Several browsers have examined methods of altering the playback mechanism in order to address this problem.

The primary technique used for achieving this is to speed up the playback rate whilst simultaneously ensuring that the pitch of the speakers remains unchanged. This process is generally implemented using an overlap and add algorithm (e.g. Hejna [1990]) which effectively has a 'concertina' like effect on the audio waveform. The concertina effect is inaudible since the 'folds' are chosen to align with pitch periods and so the technique is similar to removing a number of pitch vibrations - thus the speech is shortened but the pitch is unchanged. Arons [1997] made use of speed up in the Speech Skimmer device which allowed the user to jointly choose the playback rate and to restrict the playback to relevant segments of the recording (computed according to prosodic cues). Other work has examined the use of speed up (e.g. Tucker and Whittaker [2006a], Arons [1992]) with the general finding that speeds of up three times real time can be understood if enough

training is given. However without training sped up speech can sound disconcerting and is a long way from natural speech even with pitch correction.

An alternative technique for allowing listeners to process speech recordings is to use information retrieval and natural language processing algorithms to identify generically 'important' regions of the recording and playback only those regions. This naturally means that the listener is no longer hearing the full recording but the approach has the advantage of not requiring any training since the speech is played back at the natural rate and also that the cognitive limit on the level of compression. Different methods can be used to compute which parts of the recording are important ranging from simple IR metrics (Tucker and Whittaker [2006b]) to more complex summarisation inspired techniques (Murray and Renals [2006]). These two examples make use of speech transcripts to derive the importance scores although it should be noted that this approach to temporal compression of speech need not require transcripts since the importance scores could be computed from purely acoustical measures (Arons [1997]). In addition to this the techniques used are fairly robust to speech transcription errors and it is typically found that the portions of the recording that are scored as having a high importance are well recognised by speech transcription systems (Zechner and Waibel [2000]).

In addition to systems which manipulate the audio stream to allow users to process speech recordings more efficiently there are also approaches which alter the method of *presenting* the audio for the same purpose. An example is Schmandt and Mullins [1995] where different parts of the recording are played simultaneously to both ears. The listener is able to attend to both parts and can focus on parts of the recording that they find interesting. Using the same technology in an alternative way Schmandt [1998] describes an audio playback system where listeners travel down a virtual hallway, hearing snippets of interesting conversation. If the listener identifies a portion of the recording that they are interested in then they can enter the relevant virtual room.

Video recordings have similar problems to those seen for audio recordings. Videos are relatively easy to make and store (although this is more complex than audio) but they are costly to search and browse. Systems that focus on video recordings have largely focused on summarisation and altered playback mechanisms.

3.1.4 Keyframing

One method of overcoming the problem of navigating lengthy video recordings is to represent the video recording as a finite number of *keyframes*. There are a variety of methods for determining which frames should be used as keyframes varying from a random selection to methods which measure the uniqueness of each frame in a series and select the most unique frame to be the designated keyframe.

Typically keyframes are presented in a linear fashion which reflects the temporal evolution of the video. Girgensohm et al. [2001] presented keyframes in a comic book style display by picking keyframes and then measuring the relative importance of each of the keyframes that had been chosen. The keyframes were then laid out in a comic book style where the importance of each keyframe is used to determine how much space the keyframe should take up in the layout. Since the layout is now two dimensional though, the temporal connection is not as clear with this layout as it is with a linear layout.

3.1.5 Video Playback

Other systems have addressed the browsing and access problems by using indexing systems similar to that seen for audio. For example, He et al. [1999] describe a video skimming system which allows users to jump backwards and forwards in time using automatically derived index markers. Additionally, the system allows the user to playback the video at increased speed using the techniques described above for speeding up audio and synchronising the audio with the video.

Foote et al. [1998] also describes a video browser with a variable speed control. Here the user has the option of manually altering the playback rate using a slider mechanism. In addition to manual control the browser also offers an automatic method of varying the playback rate. The playback rate is linked to a confidence measure of 'interestingness' so that the user watches relevant portions of the recording in real time or near real time and portions of the recording which are measured as being uninteresting are played back at a much faster rate.

3.1.6 Video Summarisation

Another means of assisting browsing of digital video is through the creation of skims - an automatically derived multimedia summary of a video (Christel et al. [1998]). The skims here have similarity with the audio skims described above but a key difference is that they not only include a video component but also de-synchronize the audio and video tracks when producing the skim. Thus the skimming process selects the important audio and video sections (which may or may not coincide) and then combine these in a coherent and meaningful way.

3.2 Second Tier Browsers

Second tier browser make use of the interaction technique described in the previous tier but additionally include participant generated data. Therefore the second tier browsers make use of slides, participant notes and any other data that is shown in the meeting or created as a by product of the meeting itself. The type of data produced in meetings can be further categorized into that which is produced by individuals and that which is produced by the community of participants. Artefacts like slides and whiteboard annotations are examples of the latter category and personal notes is a good example of individual data which is produced in a meeting. We examine each of these types of data below.

3.2.1 Slides

Geyer et al. [2001] describes the TeamSpace system which includes elements to support the organisation of meetings as well as providing means to record a meeting and a corresponding interface to review archived meetings. The meeting viewer includes some of the audio segmentation and indexing work described above but centrally focuses on the slides that were presented during the meeting. Thus the user is able to select a slide and listen to audio that was said whilst the slide was being displayed. Additionally the system records any annotations made to the slide.

3.2.2 Whiteboard

The Distributed Meetings client (Cutler et al. [2002]) is another system which integrates several components into a meeting organiser and recorder. The system records video using a panoramic view of the whole meeting room and records audio using a single microphone array to aid localisation and tracking of meeting participants. The system also uses a separate digital camera to capture the whiteboard which has the advantage of capturing who is writing on the board, as well as any non-annotation gestures (such as pointing etc.). The resulting browsing interface shows the whiteboard image centrally, along with audio and video segmentation information. The whiteboard markings are also segmented, again allowing users to select a whiteboard segment and watch the audio and video related to that segment.

Brotherton et al. [1998] describe a system for the visualisation of multiple media streams for the Classroom 2000 project. The purpose of this project is to take a lecture, capture multimedia data from the presentation and then package this data together in a format that supports post hoc browsing and information extraction. The system uses a digital whiteboard to capture annotations during a lecture and then uses post-processing on the whiteboard annotations for segmentation. The level of granularity is much greater here, therefore, since the user is able to select single annotations and determine from this which specific part of audio was being played when this annotation was made. The system also provides a ‘focus of attention’ timeline which indicates at what point during the lecture slides or the whiteboard were the main focus of the class.

3.2.3 Notes

The most common method of integrating personal notes into the browsing interface is by time-stamping each individual ‘note’ and then using this as a supplementary index into the audio and video recordings. This is the approach of the Filochat system (Whittaker et al. [1994]) where a tablet PC was used by individual participants as a means of taking notes during a meeting. The PC also acted as a means of recording the meeting audio. The users of the system could then revisit their notes after the meeting, select a particular annotation and hear the audio that was recorded at the time the note was taken.

A similar approach was taken by Moran et al. [1997], although here the application was designed to be used by a single person rather than by all the participants in a meeting. The chair of the meeting used a PC to write notes in a specific template and then used these notes to revisit the meeting and make a particular decision. Both these studies found that in addition to supporting note taking practices the introduction of these system lead to changes in the way that participants take notes. Specifically, Moran et al. [1997] noted that users of the system would often make short notes (“ha” was used for this) to indicate something that was interesting and that should be re-listened to later. Chiu et al. [1999] also implements this paradigm but extends it to account for multimedia. Here users are able to annotate chosen video frames or presented slides as they wish. They can also select an automatic setting where new slides or significant changes to the video add a new pane to the display onto which the user can make notes. The user can then review the recording by looking at the static slide and video captures.

3.2.4 Minutes

The final area of participant created data are the meeting minutes. To support the taking of minutes Chiu et al. [2001] allowed the designated scribe to take the minutes on a wirelessly connected laptop. Whilst the minutes are being taken an audio, video and slide capture recording is made of the meeting. At the end of the meeting the minutes are then distributed in a variety of formats, some of which contain links back to the slides and video of relevant sections of the meeting. The minuting system also allows participants to revise the minutes as necessary, with the revisions being passed on to the other meeting participants.

The MinuteAid (Lee et al. [2004]) system takes a similar approach. The system here differs in that it allows the scribe to take minutes during the meeting and add any multimedia created during the meetings (slides, video frames) as they desire during the meeting. Thus the resulting meeting minutes are manually constructed but with user specified references to the multimedia content.

3.3 Third Tier Browsers

The third tier of meeting browsers have access to the views and data provided by the first and second tiers but embellish these with data derived from the raw meeting content and different presentations of this data. Whereas most of the browsers above were focused on single types of data or presentations the browsers in this tier tend to take a broad view and, in some cases, could be considered fully featured state of the art meeting browsers.

Whittaker et al. [2002] outlines the ScanMail system which, although is designed to work with voicemail, has functionality which could be applied to meetings. In ScanMail voicemail messages are converted into enhanced emails using a combination of speech transcription and post processing. When a voicemail is left for the user the system converts it into text using an automatic speech transcriber. The user then has a perspective on their voicemail which is like an email reader. The speech and text is synchronised so that the user is able to listen to specific parts of the voicemail by selecting the relevant parts of the text. Thus if there are any sections of the transcript which appear to contain transcription errors the user can immediately verify what the correct text should be by listening to the corresponding portion of audio. The system also allows user to search their voicemail with a text search and marks up parts of the message, such as phone numbers, so that users can easily extract important information from the message.

The Rough N Ready browser (Colbath et al. [2000]), like ScanMail, is not focused on the meeting domain but again contains ideas which are applicable to meeting browsers. The system starts by processing news recordings and transcribing the speech. Again the transcription links back to the audio recording so the user can choose to listen to a portion of the audio recording at any time by selecting a part of the transcript. Additionally users are able to search for specific entities, such as people, locations, organizations. Search results can be displayed on a timeline indicating the temporal density of the search results.

A meeting browser which gives the transcript prominence is described in Bett et al. [2000]. Here the interface consists of the transcript and a single video component alongside a list of participants and a timeline indicating when each of the participants was

speaking. In addition to this archival browser the system allows the user to construct summaries using audio, video and text of the whole meeting or specified parts of the meeting. The interface also allows for the display of various discourse features in a browser and also allows users to search the entire meeting archive.

Lisowska et al. [2004b] describe ARCHIVUS, a system designed to allow users to browse and access multimodal meetings through search or by browsing. The system uses a library metaphor in its interface. Thus each meeting is represented as a book on a shelf - opening a book from the shelf reveals the transcript of the meeting. In addition to the textual transcript the user has access to multimedia elements related to the meeting. In this system the user also has the choice of accessing the archive through speech.

A browser which also examined search is the Transcript-Based Query and browsing interface (TQB) described in Popescu-Belis and Georgescul [2006]. The TQB interface allows the user to enter free text queries and search over a set of meeting transcripts for utterances which contain these queries. The interface also allows user to focus their searches, for example by searching for utterances by a single participant or utterances which are of a particular type (e.g. a question). The TQB also allows the user to browse the meeting archive by selecting particular episodes or keywords of a particular meeting or by jumping into points where certain documents were discussed. ARCHIVUS and TQB were developed as part of the IM2 project which also developed a number of different meeting browsers which are described in detail in Lalanne et al. [2005].

Jabber (Kazman et al. [1996]) and Ferret (Wellner et al. [2004]) take a similar approach to visualising a recorded meeting. Both focus on a temporal view showing which participants spoke at which point of the meeting. Both contain video components, with Ferret allowing for multiple video components showing different views of the meeting. In addition to these components the Jabber browser shows an overview of the meeting by plotting a graph of involvement for each participant over the course of the whole meeting

The final browser in this category is the document focused browser (Lalanne et al. [2003]). Again this browser displays a transcript and segments the audio into discrete meeting sections. The browser also shows the participant involvement but uses a circular representation to show this alongside several other types of temporal metadata. The focus of this browser is, however, a document and the browser is able to highlight parts of the document that are currently being discussed. Thus the user is able to select a part of the document and then hear the audio that relates to that particular section. This notion has also been extended to look at relationships between discrete multimedia elements (Lalanne et al. [2007]).

4 State of the Art Browsers

We now compare two state of the art browsers. JFerret is an extension of the Ferret browser described above, and the Portable Meeting Recorder (Lee et al. [2002]) is a portable meeting recording and browsing system. Both are interesting because although they fall into the third tier of browsing they do more than just visualising raw data streams and offer a flavour of a possible fourth tier of browsing.

The portable meeting browser (see Fig. 3) encompasses all stages of the meeting capture process. The processing begins with a small single camera which uses a parabolic mirror to enable it to capture a full 360° display of the meeting. At the base of the camera are four microphones placed in a square formation to allow for beam forming which used in later stages of processing. Thus the recordings made are just audio and video and the recording interface allows users to make annotations as the meeting is recorded but this seems to be intended for users of the system rather than for making personal notes.

Following the meeting, five sets of post processing are carried out to produce meeting metadata. Firstly the audio streams are processed in order to localize each speaker in the meeting room, here the azimuth angles of each speaker is computed relative to the recording device. On the basis of this an algorithm produces a single video stream which automatically determines the best view for the meeting recording. The current speaker location and a measure of visual change is used to identify the best frame in each case. The background image is then extracted and matched against a database of room templates in order to identify the meeting location.

Once these automatic annotations have been made and placed into the metadata database, the meeting description document is produced which contains all the information that has been automatically extracted - the meeting date, time, location and participants along with an image of the participant. The user can then access the archived recording and search and automatically produced speech transcription to locate points of interest and then watch the corresponding portions of the meeting. In addition to the text search the interface offers the option for users to scan a graph of audio and visual activity in order to locate points of interest.

All of these data are then placed into a rich user interface which includes the activity indexes, speaker participation, key frames of the video, the chosen best shot, the overall panoramic view and the meeting overview. Users can jump to any point in the meeting by selecting from these components

JFerret (see Fig. 2) differs from the browsers above (including the Ferret browser) in that it is a *framework* for building meeting browsers (e.g. Murray et al. [2007]). JFerret allows browser designers to not only layout the various components in a way that suits the intended application but also the components themselves can be altered and chosen in order to make the most efficient browser for the given application. The user of JFerret selects which components they wish to include in their browser and uses a simple XML file to specify where these components should be placed in the screen display. The framework ensures offers a central point of synchronisation which ensures that changes made in one component are reflected in all of the other components in the display.

Thus the expressive power of JFerret can be found in the superset of all the components that are available to the browser designer. In keeping with the browsers described above JFerret offers a component for displaying video and playing back audio. The system also is able to show slides, any kind of speech transcript, and the personal notes of each of the meeting participants. In addition to these more basic components JFerret includes components for playing back audio at faster rates using the overlap and add techniques described above and more semantically motivated playback techniques (Tucker and Whittaker [2006a]). The browser can also include a summarisation component consisting of

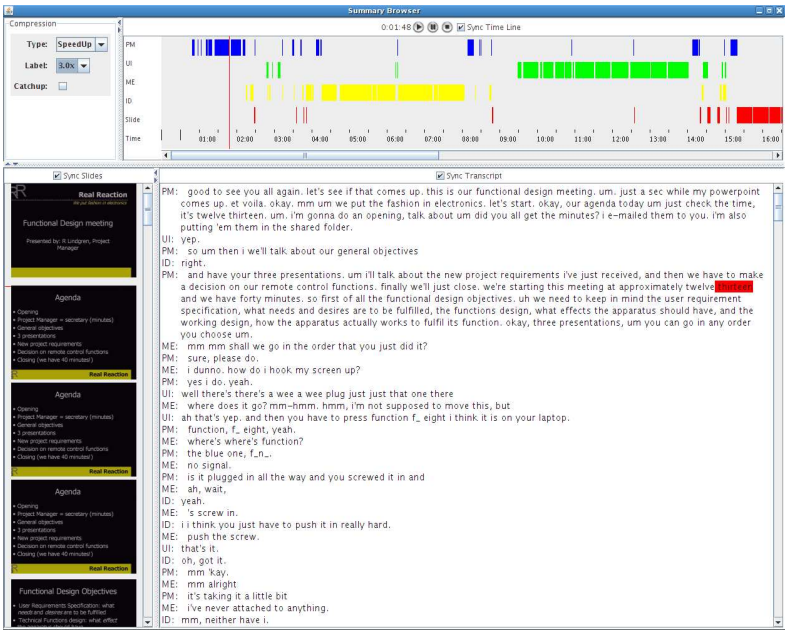


Figure 2: The JFerret Meeting Browser showing a time line, compressed audio player, meeting slides and transcript.

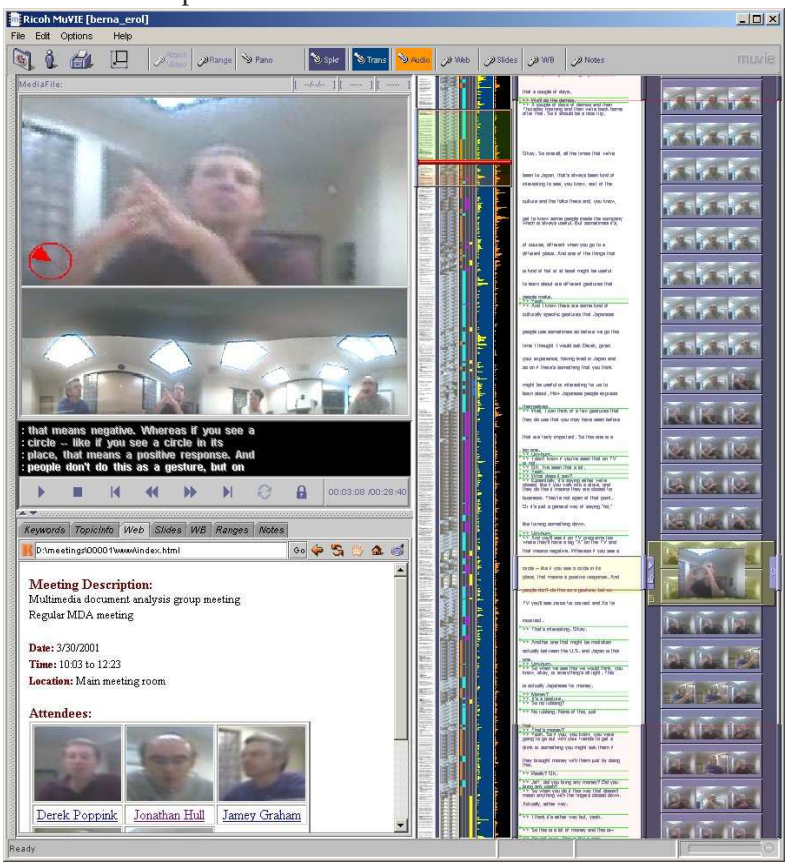


Figure 3: The Portable Meeting Browser showing videos, keyframes, transcripts, overviews and activity graphs.

both extractive and abstractive summaries which are linked back into the transcript so that the user can select a part of the summary and see where in the transcript the phrases originated.

The display can also include components which indicate the dominance of each speaker (Rienks [2007]), allowing the user of the browser to identify who is the dominant speaker at any point during the meeting. In addition to this there is also a component which allows the user to annotate and track the strands of various arguments throughout the meeting. Arguments are shown in a tree like structure showing each thread of the argument and when parties are in disagreement.

JFerret also allows the user to directly search the ASR output in order to locate keywords (Szoke et al. [2005]). This search is carried out on the ASR lattice and is, therefore, more powerful than doing a text search of the meeting transcript as potential candidates for the keyword match which were rejected by the ASR algorithm can be included in the search. In addition to this users are also able to do a simple text search over the meeting transcript. In addition to these components JFerret also includes a device to allow people to share three dimensional data models and manipulate them whilst in a meeting.

5 Evaluation

Given that meeting browser is still an emergent field it is unsurprising that little work has addressed the evaluation of the efficiency of the meeting browser (see Cremers et al. [2006] for a summary of AMI work). However, now that standard corpora exist Carletta [2006] future work will address this problem. Two relatively large scale evaluations have taken place which suggest directions that future evaluations will take.

The Meeting Browser Evaluation Test (BET) takes a TREC[1] like approach to evaluating meeting browsers. There is a lengthy one-time data collection process and then a relatively rapid evaluation between subjects evaluation phase. The advantage of this process is that the results of the data collection phase can be shared between evaluators and so evaluations can take place in different locations and at different times.

The core of the BET is the notion of *observations of interest*. An observation of interest is pair of statements (one true and one false) which addresses a singular fact about the meeting. Thus an observation may be “The budget was 100 pounds” paired with “The budget was 300 pounds”. In the data collection phase a small number of judges watch a meeting in its entirety and then re-watch the meeting and generate these observations of interest. The collection of observations of interest form the basis of the test set which can be shared between evaluators. In the evaluation phase, users are given a novel meeting browser and asked to validate as many observations of interest as they can - browser efficiency is then measured as the number of observations that can be validated in a given amount of time. The experimenters can also log the media time points that the observations were answered at given an indication of whether the user was guessing the answer or actually spotted it in the meeting. The BET framework has also been extended to allow experimenters to have more control over what kind of observations are used when evaluating browsers and this

1. http://trec.nist.gov/

extended form of the BET has been used for evaluation campaigns Popescu-Belis et al. [2007].

It could be argued that the BET is an intrinsic evaluation - it measures browser performance in a simulated task that approximates one use of the browser in the real world. An example of an extrinsic evaluation can be found in Elling [2007]. Here browser performance is measured by adding the technology into a simulated meeting and seeing how the meeting process is improved as a result of supplying different browser systems to the team. The users are told that they are replacing a team who have previously met regarding a project to build a new remote control for the television. The previous meetings were recorded and the new team are provided with different meeting browsers in order to review the prior work. A large number of performance measures are then used to measure how successful each team is.

The extrinsic evaluation has the advantage of placing the users in a more realistic environment where the performance of the browser is critical to their success. The drawback to this is that the experimenter is unable to make fine-grained assessments of the browser performance - it is difficult to draw out why a particular browser worked well or what specific questions the browser was adept at answering. These types of questions could be answered by a BET style evaluation approach.

6 Conclusion

We have shown how meeting browsers can be considered to be in one of three tiers depending on the type of data that the browser focuses on. First tier meeting browser make direct use of the raw data streams that are recorded during a meeting - thus they concentrate on the audio or video. Second tier browsers make use of and focus on the data that the meeting participants create or present during the meeting - slides, minutes, personal notes etc. Third tier data is that derived from the raw and participant data namely things like ASR transcripts, participant involvements, locations etc. Browsers in each tier generally make use of principles and data from the tiers above it so that third tier browsers can be considered state of the art browsers. We also examined two third tier browsers in detail - one which was a complete meeting recording and browsing solution and one which is effectively a framework for combining browser components into a single browser. We also briefly examined how such browsers are evaluated.

References

B. Arons. Techniques, perception, and applications of time-compressed speech. In *1992 Conference, American Voice I/O Society*, pages 169–177, September 1992.

B. Arons. Speechskimmer: A system for interactively skimming recorded speech. *ACM Transcations on Computer-Human Interaction*, 4(1):3–38, March 1997.

M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel. Multimodal meeting tracker. In *RIAO*, April 2000.

M. Bouamrane and S. Luz. Meeting browsing: state-of-the-art review. *Mulitmedia Systems*, 12:439–457, 2007.

J. A. Brotherton, J. R. Bhalodia, and G. D. Abowd. Automated capture, integration and visualization of multiple media streams. In *The IEEE International Conference on Multimedia Computing And Systems*, pages 54–63, 1998.

J. Carletta. Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5, 2006.

P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox. Notelook: Taking notes in meetings with digital video and ink. In *ACM Multimedia '99*, 1999.

P. Chiu, J. Boreczky, A. Girgensohn, and D. Kimber. Liteminutes: An internet-based system for multimedia meeting minutes. In *10th WWW Conference*, pages 140–149, May 2001.

M.G. Christel, M.A. Smith, C. Roy Taylor, and D.B. Winkler. Evolving video skims into useful multimedia abstractions. In *CHI '98*, April 1998.

S. Colbath, F. Kubala, D. Liu, and A. Srivastava. Spoken documents: Creating searchable archives from continuous audio. In *33rd Hawaii International Conference On System Sciences*, 2000.

A. Cremers, W. Post, E. Elling, B. van Dijk, B. van derWal, J. Carletta, M. Flynn, P. Wellner, and S. Tucker. Meeting browser evaluation report. Technical report, AMI Project Deliverable, 2006.

R. Cutler, Y. Rui, A. Gupta, J.J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed meetings: A meeting capture and broadcasting system. In *10th ACM International Conference on Multimedia*, pages 503–512, December 2002.

L. Degen, R. Mander, and G. Salomon. Working with audio: Integrating personal tape recorders and desktop computers. In *CHI '92*, pages 413–418, May 1992.

E. Elling. Tools for fun and fruitful meetings. Master's thesis, University of Twente, 2007.

J. Foote, J. Boreczky, A. Girgensohn, and L. Wilcox. An intelligent media browser using automatic multimodal analysis. In *ACM Multimedia*, pages 375–380, September 1998.

W. Geyer, H. Richter, L. Fuchs, T. Frauenhofer, S. Daijavad, and S. Poltrock. A team collaboration space supporting capture and access of virtual meetings. In *2001 International ACM SIGGROUP Conference On Supporting Group Work*, pages 188–196, September-October 2001.

A. Girgensohm, J. Borczky, and L. Wilcox. Keyframe-based user interfaces for digital video. *IEEE Computer*, 34(9):61–67, September 2001.

L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-summarization of audio-video presentations. In *7th ACM International Conference On Multimedia*, pages 489–498, 1999.

D.J. Hejna. Real-time time-scale modification of speech via the synchronized overlap-add algorithm. Master's thesis, M.I.T., 1990.

D. Hindus and C. Schmandt. Ubiquitous audio: Capturing spontaneous collaboration. In *1992 ACM Conference on Computer-Supported Cooperative Work*, pages 210–217, November 1992.

A. Jaimes, K. Omura, T. Nagamine, and K. Hirata. Memory cues for meeting video retrieval. In *CARPE '04*, pages 74–85, October 2004.

R. Kazman, R. Al-Halimi, W. Hunt, and M. Mantei. Four paradigms for indexing video conferences. *IEEE Multimedia*, 3(1):63–73, Spring 1996.

D. Lalanne, S. Sire, R. Ingold, A. Behera, D. Mekhaldi, and D. Rotz. A research agenda for assessing the utility of document annotations in multimedia databases of meeting recordings. In *3rd International Workshop on Multimedia Data And Document Engineering*, September 8th 2003.

D. Lalanne, A. Lisowska, E. Bruno, M. Flynn, M. Gerogescul, M. Guillemot, B. Janvier, S. Marchand-Maillet, M. Melichar, N. Noenne-Loccoz, A. Popescu-Belis, M. Rajman, M. Rigamonti, D. Rotz, and P. Wellner. The IM2 multimodal meeting browser family. Technical report, IM2 Project, 2005.

D. Lalanne, M. Rigamonti, F. Evequoz, B. Dumas, and R. Ingold. An ego-centric and tangible approach to meeting indexing and browsing. In Bourlard H. & Renals S. Popescu-Belis A., editor, *Machine Learning for Multimodal Interaction IV, Revised Selected Papers, LNCS*. Springer-Verlag, Berlin/Heidelberg, 2007.

D. Lee, J.J. Hull, B. Erol, and J. Graham. Minuteaid: Multimedia note-taking in an intelligent meeting room. In *IEEE International Conference on Multimedia and Expo*, 2004.

D-S Lee, B. Erol, J. Graham, J. J. Hull, and N. Murata. Portable meeting recorder. In *ACM Multimedia*, pages 493–502, 2002.

A. Lisowska. Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study. Technical Report IM2.MDM Report 11, IM2, November 2003.

A. Lisowska, A. Popescu-Belis, and S. Armstrong. User query analysis for the specification and evaluation of a dialogue processing and retreival system. In *Proceedings of LREC 2004*, volume III, pages 993–996, 2004a.

A. Lisowska, M. Rajman, and T.H. Bui. Archivus: A system for accessing the content of recorded multimodal meetings. In *MLMI 2004*, 2004b.

T.P. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. Melle, and P. Zellweger. "i'll get that off the audio": A case study of salvaging multimedia meeting records. In *CHI '97*, 22-27 March 1997.

G. Murray and S. Renals. Dialogue act compression via pitch contour preservation. In *Proceedings of the 9th International Conference on Spoken Language Processing, Pittsburgh, USA*, September 2006.

G. Murray, P. Hsueh, S. Tucker, J. Kilgour, J. Carletta, J.D. Moore, and S. Renals. Automatic segmentation and summarization of meeting speech. In *Proceedings of NAACL-HLT 2007*, April 2007.

A. Popescu-Belis and M. Georgescul. Tqb: Accessing multimedia data using a transcript-based query and browsing interface. In *Proceedings of LREC 2006*, pages 1560–1565, 2006.

A. Popescu-Belis, P. Baudrion, M. Flynn, and P. Wellner. Towards an objective test for meeting browsers: the BET4TQB pilot experiment. In Bourlard H. & Renals S. Popescu-Belis A., editor, *Machine Learning for Multimodal Interaction IV*, pages 108–119. Springer-Verlag, Berlin/Heidelberg, 2007.

R. Rienks. *Meetings in smart enviornments: Implications of progressing technology*. PhD thesis, University of Twente, 2007.

C. Schmandt. Audio hallway: A virtual acoustic environment for browsing. In *UIST*, pages 163–170, 1998.

C. Schmandt and A. Mullins. Audiostreamer: Exploting simultaneity for listening. *Proceedings of CHI '95*, 1995.

I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, and J. Cernocky. Phoneme based acoustics keyword spotting in informatl continuous speech. In V. Matousek, editor, *Lecture Notes In Computer Science*, volume 2658, pages 302–309. Springer-Verlag, 2005.

S. Tucker and S. Whittaker. Accessing multimodal meeting data: Systems, problems and possibilities. In S. Bengio and H. Bourlard, editors, *Lecture Notes In Computer Science*, volume 3361, pages 1–11. Springer-Verlag, 2005.

S. Tucker and S. Whittaker. Displaying dynamic meeting transcripts: Concertina browsing. In *Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, May 2006a.

S. Tucker and S. Whittaker. Time is of the essence: An evaluation of temporal compression algorithms. In *Proceedings of CHI '06*, April 2006b.

P. Wellner, M. Flynn, and M. Guillemot. Browsing recording of multi-party interactions in ambient intelligent envrionments. In *CHI*, April 2004.

S. Whittaker, P. Hyland, and M. Wiley. Filochat: Handwritten notes provide access to recorded conversations. In *Chi '94*, 271-277, April 1994.

S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, P. Isenhour, L. Stead, G. Zamchick, and A. Rosenberg. SCANMail: A voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of CHI 2002*, April 2002.

S. Whittaker, S. Tucker, K. Swampillai, and R. Laban. Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing*, 2008. In press.

K. Zechner and A. Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL-200*, pages 186–193, Seattle, WA., April / May 2000.