Using Speech-Specific Characteristics for Automatic Speech Summarization

Gabriel Murray

Doctor of Philosophy

Centre for Speech Technology Research
Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh

2008

Abstract

In this thesis we address the challenge of automatically summarizing spontaneous, multi-party spoken dialogues. The experimental hypothesis is that it is advantageous when summarizing such meeting speech to exploit a variety of speech-specific characteristics, rather than simply treating the task as text summarization with a noisy transcript. We begin by investigating which term-weighting metrics are effective for summarization of meeting speech, with the inclusion of two novel metrics designed specifically for multi-party dialogues. We then provide an in-depth analysis of useful multi-modal features for summarization, including lexical, prosodic, speaker, and structural features. A particular type of speech-specific information we explore is the presence of meta comments in meeting speech, which can be exploited to make extractive summaries more high-level and increasingly abstractive in quality. We conduct our experiments on the AMI and ICSI meeting corpora, illustrating how informative utterances can be realized in contrasting ways in differing domains of meeting speech. Our central summarization evaluation is a large-scale extrinsic task, a *decision audit* evaluation. In this evaluation, we explicitly compare the usefulness of extractive summaries to gold-standard abstracts and a baseline keyword condition for navigating through a large amount of meeting data in order to satisfy a complex information need.

Acknowledgements

I would like to thank my colleagues at the Centre for Speech Technology Research (CSTR), the Human Communication Research Centre (HCRC), the Institute for Communicating and Collaborative Systems (ICCS) and the Department of Linguistics and English Language at the University of Edinburgh, for stimulating conversations, a great cooperative and inter-disciplinary work environment, and the numerous personal friendships that grew out of working relationships. I feel very fortunate to have had the opportunity to pursue my postgraduate work at the University of Edinburgh, where there are many channels of communication and collaboration between schools, departments and research centres. Being part of the community of speech and language researchers here has been fantastic, invaluable and unforgettable.

I owe a particular debt of gratitude to Steve Renals, my advisor during my MSc. and PhD. programs. Your advice and guidance has always been illuminating, incisive and thought-provoking, and I've benefited hugely from researching with you. I can't imagine a better academic scenario than spending the past few years under your counsel. I often came away from our weekly meetings envisioning a given task in an entirely new way or suddenly spotting a new route to a solution as a result of our conversations. There has always been the perfect balance between guidance and independence, and I've really appreciated that.

Many thanks also go to Simon King, who was the course organizer for the MSc. in Speech and Language Processing when I took that course. Simon, you have an exceedingly rare gift for teaching this material in a clear, accessible way, and I hope that many more students have the opportunity to learn it from you.

My PhD. research was partly funded by the AMI and AMIDA projects, and I would like to thank all of my AMI Consortium colleagues for the opportunity to work and share with you. It's been great to be a part of this team and to see the variety of interesting research happening under the AMI umbrella. I would particularly like to thank the project leaders, Steve Renals and Hervé Bourlard, as well as WP5 manager Tilman Becker. Thanks also to the Edinburgh WP5 team: Steve Renals, Johanna Moore, Jean Carletta, Pei-Yun Hsueh, Alfred Dielmann, Weiqun Xu, Jonathan Kilgour, Theresa Wilson and John Niekrasz. A special thanks also to Thomas Kleinbauer, who was researching abstractive summarization at DFKI while I was researching extractive summarization at Edinburgh, for numerous enjoyable conversations about summarization directions. A big thanks to the AMI-ASR team, and to Thomas Hain and Giulia Garau in particular for helpful discussions.

Many thanks to Mark Steedman and Julia Hirschberg for your valuable feedback and thoughtful comments on an earlier draft. The thesis has benefited greatly from your review and suggestions, and I very much appreciate your time and consideration.

I would also like to thank the Overseas Research Scheme (ORS) and the university's matching scheme for providing funding, which allowed me to stay in Edinburgh after my MSc. had been completed. I am very thankful to have had these years of research in Edinburgh and it wouldn't have been possible without this generous allowance.

Sofie, it was very cool to be finishing our PhDs around the same time and to be able to send each other brief daily emails of support. It'll be fantastic to sit together over two foamy cappuccinos and realize that we're done. Thank you for your friendship and encouragement, and for the treasured visits that punctuated our increasingly hectic schedules these past years.

Thank you so much to all of my family for your support and encouragement during these years. You've all helped me so much, in ways large and small, and I wouldn't have been able to do it without you. At times it's been difficult being far away from home for such a long stretch and only being able to catch up in person once or twice a year, and I'm thrilled to be able to move closer to you all now and have the freedom to drop in for a visit on a whim.

Heather, my biggest thanks is to you, for all your love and support. During the stress of the PhD., I could always count on you for encouragement, laughter, and spontaneous dance parties in our flat. By far the greatest thing about my postgraduate studies is that they brought me to Edinburgh, where you and I met in circumstances that are hard to believe and which changed my life so completely and wonderfully. I'm so happy and thankful as we make this new transition in our lives. This thesis is dedicated to you for your support and encouragement as I wrote it, your belief in me, your companionship, and the way you inspire me through your kindness, intelligence and unsurpassed goofiness.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Gabriel Murray*)

To Heather

Table of Contents

Chapter 1

Introduction

Speech summarization is the process of digesting speech data and presenting only the most informative or most relevant source information, thereby providing a distilled version of the source as a substitute for, or an index into, the original. In the research described herein, the input consists of spontaneous multi-party speech and the summarization process results in automatically generated overviews of meeting discussions, analogous to human minutes of a meeting.

While the field of text summarization has grown steadily over recent decades, speech summarization is comparably young and under-developed. Robust algorithms have been developed for summarizing text data such as news-wire and articles, and annual summarization challenges such as the Document Understanding Conference (DUC)[1] chart the continuing progress of the text summarization community. By and large, methods for summarizing various forms of speech data are yet to be fully explored and evaluated. One of the aims of this work is to examine how advances in text summarization might be applied to the domain of speech summarization; while speech data presents a more complex summarization challenge than relatively well-formed text data, knowledge transfer between the two overlapping summarization communities should be of benefit to all. And with speech summarization being the younger of the two fields, it seems most sensible for speech summarization researchers to begin their exploration by applying proven textual approaches to the speech data at hand.

The second, much larger theme of this work is the search for useful speech-specific characteristics in automatic speech summarization. While it is desirable to exploit text summarization advances as much as possible, the unique nature of speech suggests that there will be features particular to the data indicating salience and relevance for

[1] http://duc.nist.gov

our purposes. Compared with purely textual data, spontaneous speech has many levels of information to investigate for our purposes, from prosodic features to turn-taking and dominance relations to unique structural features of meeting speech. The central hypothesis of this paper is that it is advantageous to include these features in any speech summarization system and that a solely text-based approach will tend to be less robust than a system incorporating these "extra" sources of information.

The form of summarization used herein is of an *extractive* variety, in which important sentences - or *dialogue acts*, in our case - are extracted and concatenated to form a summary comprised of important bits of the meeting. This is quite different from the popular human conception of a summary, wherein novel sentences are created to briefly convey the information content of the source material. Thus, while our resulting summaries are analogous to human minutes of a meeting, they are distinct in form. The advantages of choosing such a summarization paradigm, however, are that extractive summarization techniques do not require a deep understanding of the source material, the techniques are relatively robust to disfluent, fragmented speech, and extractive summarization methods are also largely domain independent. In contrast, abstractive summarization normally requires a deeper understanding of the source material, a method of transforming the source representation to a summary representation, and a natural language generation component to create novel summary sentences. While the summarization work described here is firmly in the extractive tradition, one theme of this research is finding out how to move summarization further down the extractive-abstractive continuum and essentially make extractive summaries more intelligent by exploiting information beyond simple binary labels of "informative" and "uninformative" and to incorporate as much high-level perspective in the summaries as possible.

Though extractive summaries will still tend to be less readable than human abstracts typically are, due to the fact that they are comprised of units that have been removed from their original contexts, it is also important to stress that these extractive summaries are not simply stand-alone textual documents. They are meant to serve as aids to the navigation of meeting content in the context of a multi-media meeting browser. In this thesis, we create an extrinsic evaluation that tests the usefulness of such summaries in aiding a real-world information-gathering task. The hypothesis for that extrinsic evaluation is that extractive summaries provide a more efficient way of navigating meeting content than simply reading through the transcript and using the audio-video record, or navigating via keyword search. This hypothesis is related to the prevalence of meeting browser use-cases that involve time-restricted users. Few people

have the time or desire to review a meeting by listening to or reading everything that was said. The extractive summaries created here are intended to enhance the experience of reviewing and attending meetings by presenting the meeting information in a condensed form to the user, and allowing the user to treat that condensed information as a spring-board to further navigation of the meeting content.

There are numerous reasons for choosing to investigate automatic summarization on meeting speech rather than other speech domains such as Broadcast News, lectures or telephone speech. The first is that meeting speech is normally purely spontaneous speech, with no read or planned portions (some meetings may contain semi-planned speech, such as brief slide presentations). The meeting corpora therefore exhibit the full spectrum of disfluencies that characterize "real" human speech. The second reason is that such corpora are comprised entirely of multi-party speech, featuring complex personal interactions, speaker overlaps, differences in speaker status, and information that may be spread across several speakers. Third, meetings are a ubiquitous part of life for many people, and technologies that enhance the meeting experience and allow meeting participants to become more efficient both during and between meetings are generally beneficial in the real world.

Meeting data are particularly interesting because the interactions are often multi-modal, featuring not just spoken dialogues, but also note-taking, emails, slides, white-board events and interactions with remote participants. Meetings are structured along all of these lines and the interactions are complex, thereby yielding numerous multi-modal features for potential exploitation in summarization research. This multi-modal aspect of the data also leaves the possibility of having summary output that is not strictly text or audio; the summaries themselves can be multi-modal in nature. This work is done as part of the Augmented Multi-party Interaction (AMI) and Augmented Multi-party Interaction With Distance Access AMIDA projects[2], which aim to develop technologies that both exploit and enhance the multi-modal aspects of meeting speech.

There is also a challenge in that automatic speech recognition (ASR) on this data is imperfect, and the word error rates (WER) tend to be much higher than you would find with domains such as broadcast news. Throughout this thesis we assess the ramifications on the summarization task of using considerably noisy data.

[2]http://www.amiproject.org

1.1 Thesis Overview

Before proceeding to a description of the core research, Chapter 2 (page 7) provides an overview of previous summarization work on text and speech data, contains a discussion of evaluation techniques, and places our automatically generated summaries in a summarization typology. Chapter 3 (page 25) gives an overview of the data used and the general experimental overview.

As mentioned above, the central hypothesis of this thesis is that for extractive summarization of spontaneous multi-party, multi-modal spoken interactions, it is advantageous to exploit a wide variety of features in the data, particularly prosodic, structural and speaker features, rather than to approach the problem on a solely textual, linguistic level. We test this hypothesis at multiple points in the summarization pipeline, as described below.

There are four major contributions of this research. First, we present results indicating which term-weighting metrics are effective for summarizing multi-party spoken dialogues, based on experiments with multiple corpora. In Chapter 4 (page 38) we describe our research comparing established term-weighting metrics from text summarization and information retrieval to novel speech-based term-weighting metrics. This research aims to establish whether there are characteristics in the speech data that can be exploited for term-weighting with the purpose of summarization. Two novel speech-based metrics are described in detail, and compared alongside more familiar text-based weighting schemes. Term-weighting can be seen as one of the first steps in the summarization pipeline, and determining an optimal term-weighting method therefore has great ramifications for all downstream processes. This will inform future speech summarization research on this type of data, as there are numerous weighting schemes to choose from and this thesis contains the first large-scale evaluation of such metrics for this data.

Second, we provide an in-depth evaluation of which features and feature subsets are effective indicators of informativeness for extractive summarization, as well as comparing unsupervised, text-based summarization approaches with supervised techniques incorporating a variety of multi-modal features. In Chapter 5 (page 63) we present several unsupervised text-based techniques and apply them to both manual and ASR transcripts for the AMI and ICSI corpora. We then present an in-depth investigation of supervised, feature-based techniques for automatic extraction for this data, building databases of lexical, prosodic, structural, and speaker features and determining

the most useful individual features and feature subsets for the extraction classification task. We also analyze how the effectiveness of given feature sets can increase or decrease when using a database aligned with manual transcripts versus ASR transcripts. The differences between summarization results on the AMI and ICSI corpora are also examined and discussed. The summarization systems in Chapter 5 are evaluated using weighted precision, recall and f-score, a novel summarization evaluation paradigm relying on multiple human annotations of extraction.

Third, we present a large-scale extrinsic evaluation for speech summarization, the *decision audit* task. While the systems described in Chapters 4 and 5 rely on the intrinsic weighted f-score metric for evaluation, Chapter 6 (page 93) describes a large-scale *extrinsic*, task-based evaluation for summarization. As mentioned above, these summaries are not meant to be stand-alone documents but rather efficient tools for browsing meetings. For that reason we would like to evaluate their usefulness in a real-world situation incorporating complex information needs. We therefore implement and describe the *decision audit* task, wherein a user must evaluate several archived meetings in order to determine why a particular decision was made by the meeting group. Incorporating five experimental conditions in total, we compare several automatic summarization approaches to gold-standard human summarization and a baseline keywords approach. We also examine the level of difficulty that ASR errors pose for time-constrained users searching for specific information. Chapter 6 (page 93) as a whole attempts to justify the extractive summarization paradigm as applied to this data, based on multiple evaluations of usability as gauged by eliciting user preferences, examining browsing behaviour and conducting human evaluations of decision audit responses. This evaluation yields very compelling results concerning the effectiveness of the extractive paradigm for multi-modal browsing of meetings, and establishes a framework for future speech summarization evaluations.

Fourth, we lay critical groundwork for moving the state-of-the-art in speech summarization further down the extractive-abstractive continuum. We recognize that extractive summarization is limited by the fact that dialogue acts lose a good deal of coherence when removed from their original contexts and that summaries comprised of utterances from within the meeting do not always offer sufficient perspective on what transpired in the discussion. While recognizing that full-scale abstractive summarization remains a lofty goal, in Chapter 7 (page 126) we lay groundwork for that ultimate objective by analyzing how dialogue acts within meetings vary between low-level and high-level perspective, and how exploiting the latter dialogue act types can

improve summarization. The phenomenon of speakers referring to the discussion itself is an informative and valuable characteristic of such data for summarization purposes. High-level informative dialogue acts are used to create "meta" summaries, which we evaluate in a number of ways. We also conduct an in-depth features analysis, describing the differing feature correlates of these distinct dialogue act types. It is hoped that this research will provide direction for moving beyond simple extraction and the reliance on strictly binary labelling of "informative" versus "uninformative." While the statistical models and features used in automatic summarization have become more sophisticated over time, it is still the case that most summarization work relies on this vague binary distinction rather than exploiting more complex distinctions in order to create more intelligent summaries.

In Chapter 8 (page 148) we discuss further work and a set of initial experiments regarding dialogue act compression, online summarization and spurt-based summarization. That chapter briefly discusses topics that may be of increased interest in the coming years and how their inherent challenges might be addressed.

Finally, Chapter 9 (page 172) concludes by giving a general overview of the results and discussing the ramifications for future summarization work on spontaneous speech data.

Chapter 2

Automatic Summarization Literature Review

2.1 Introduction

In this chapter we given an overview of the state of the art in automatic summarization. We first present a typology for summarization, and secondly review major work that has been carried out to date. We examine work on text and speech data in turn, and conclude with a review of approaches to summarization evaluation.

2.2 Types of Summaries

As mentioned in the introduction, one possible division of 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 concerns the generation 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 merely 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 (Borko & Bernier, 1975). 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 (e.g. multiple newswire articles or meeting transcripts) 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. A central focus of the AMIDA project is automated content linking for multiple meetings.

Similarly, this work focuses on *generic* summaries rather than *query-dependant* summaries. 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. The generic summarization work described herein could be extended to query-dependent summarization by combining the features of general informativeness with further measures of query overlap and responsiveness.

It is possible to divide between *text* and *speech* summarization, or *text* and *multimedia* 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 it 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.

2.3 Related Summarization Work

2.3.1 Text Summarization

Among the earliest work on automatic text summarization was the research by Luhn (1958), 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 (1969) 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 "negative relevance": 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 than with manually-tuned weights as Edmundson employed.

The ADAM system of the 1970s (Rush et al., 1971; Mathis, 1972; Pollock & Zamora, 1975) relies heavily on cue phrases, but also strives to maximize coherence by analyzing whether a candidate sentence contained anaphoric references (Endres-Niggemeyer, 1998). In the case that a candidate does contain anaphoric references, the system tries 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 are possible, the candidate is not chosen.

In the late 1970s and early 1980s, Paice (1980) investigated the idea of using "self-indicating phrases" to detect informative sentences from journal papers. These phrases explicitly signal that a sentence is relevant to the document as a whole, e.g. "This report concerns...". Contemporary work by Janos (1979) divided documents into "metatext" and "the text proper". Janos found that while most metatext could be discarded in

the summarization process, certain *thematical* metatext sentences were able to form a "semantic nucleus" for the summary as a whole. The work of both Paice and Janos has some similarity with our work in Chapter 7 (page 126) on detecting meta comments in meeting speech.

The summarization work of Paice is also similar to the ADAM summarization system in its treatment of *exophoric* sentences. The strategies are much the same: try to extract both linked sentences, else neutralize the exophoric expression, and as a last resort discount the candidate sentence. The primary difference is that Paice evaluated both anaphoric and cataphoric references.

In the 1980s, several summarization methods arose that were inspired by findings in psychology and cognitive science (DeJong, 1982; Fum et al., 1982; Jacobs & Rau, 1990). 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 (DeJong, 1982) 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. These approaches are limited by being very domain-specific and requiring prior knowledge about the data being summarized. Further information on such approaches can be found in (Endres-Niggemeyer, 1998).

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 news-wire 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 more 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 (1969) 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 (Kupiec et al., 1995; Teufel & Moens, 1997), rather than manually tuning the weights of these features as in the work of Edmundson. For the tasks of summarizing engineering papers (Kupiec et al., 1995) and computational linguistics papers (Teufel & Moens, 1997), 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) (Mann & Thompson, 1988). 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 (Ono et al., 1994; Marcu, 1995, 1997).

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 (1997) lexical chains were detected according to the relatedness of document terms, and sentences corresponding to the strongest chains were extracted. The SUMMARIST system (Hovy & Lin, 1999) utilizes 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 (1998) introduced the Maximal Marginal Relevance (MMR) algorithm, which scores a candidate sentence according to how relevant it is to a query (or how generally relevant, for a generic summary) and how similar it is to sentences that have already been extracted. The latter score 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 (Zhu & Penn, 2006). Work by Radev et al. (2000, 2001) 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 problem 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 (1995) 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, refuelling, 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 took 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 refuelling and supplies.

Since 2001, the Document Understanding Conference 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 news-wire 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 (Lacatusu et al., 2005; Hovy et al., 2005).

Automatic text summarization is closely intertwined with automatic text retrieval, and this connection can especially be seen in query-dependent summarization, wherein a query and a document or set of documents must be represented in such a way that similarity between the query and a candidate document or sub-document can be gauged. A major difference between the tasks of text retrieval and query-dependent summarization is that text retrieval in its basic form concerns the determination of whether or not a document is relevant to a query, whereas summarization goes a step further and condenses the relevant documents. The basic formulation of the text retrieval task is that there is an archive of documents, a user who generates a query, and a process of retrieving the documents in the archive that satisfy the query's information need (Rijsbergen, 1979). An efficient way of representing queries and documents is via a vector-space representation where words are associated with term-weights, with an example weighting scheme being *tf.idf* (Jones, 1972; Rijsbergen, 1979; Salton & Buckley, 1988), where a word has a high score if it occurs often in the candidate document but rarely across the set of documents. Chapter 4 (page 38) analyzes the

term-weighting problem as applied to spontaneous speech data. The vector-space representation is useful because if both the query and candidate document are represented as vectors, similarity can be easily gauged using the cosine of the two vectors. Alternatively, probabilistic information retrieval systems (Maron & Kuhns, 1960; Rijsbergen, 1979) estimate the probability of relevance for a document D, $P(R|D)$. This is arrived at using Bayes theorem, with probability $P(D|R)$ equal to the product of the individual term probabilities in the simplest formulation (Singhal, 2001)

$$P(D|R) = \prod_{t_i \in Q,D} P(t_i|R) \cdot \prod_{t_j \in Q,\bar{D}} (1 - P(t_j|R))$$

where t_i is a term common to the query and the document and term t_j is a term present in the query but missing from the document. Since realistically the relevance information is not known, there are numerous methods for estimating the probability of a term given the relevance information, and Croft and Harper (1979) illustrate an estimation method that is closely approximated by inverse document frequency (Jones, 1972), discussed in more detail in Chapter 4.

Automated information retrieval as a field took root in the 1940s with the germinal work of Bush (1945), and it was Luhn (1958), mentioned above, who put forth the idea that words could act as indices for documents in a collection. Probabilistic information retrieval was developed in the early 1960s (Maron & Kuhns, 1960), and further refined in the 1970s and 80s (Jones, 1972; Croft & Harper, 1979). Since the early 1990s, the Text Retrieval Conference (TREC) (Harman, 1992) has encouraged the development of effective retrieval methods for large corpora (Singhal, 2001). An overview of information retrieval as a whole is outside of the scope of this thesis, but standard introductions to the field are by Rijsbergen (1979) and Salton and McGill (1983), with Singhal (2001) providing a very concise overview.

2.3.2 From Text to Speech

McKeown et al. (2005) 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. For meetings in particular, their research involved finding the prosodic and lexical corre-

lates of topic shifts, and they investigated known useful features of monologue speech such as pauses and cue phrases and concluded that these are informative for segmenting multi-party dialogue speech as well.

Christensen et al. (2003) investigated how well text summarization techniques for news-wire 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 (WER) only caused slight degradation according to their automatic metrics, but that human judges rated the error-filled summaries much more severely.

In the following sections we first provide an overview of interesting early research on speech summarization, then describe speech summarization research from four particular domains: newscasts, meetings, lectures, and voicemail.

2.3.3 Speech Summarization

In the early 1990s, simultaneous with the development of improved automatic speech recognition, researchers became increasingly interested in the task of automatically summarizing speech data. Here we describe several early summarization projects from a variety of speech domains.

Chen and Withgott (1992) 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 (1992) 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.

One of the early projects on speech summarization was VERBMOBIL (Reithinger et al., 2000), a speech-to-speech translation system for the domain of travel planning. The system is 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 can 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 is able to incorporate abstractive summarization is due to the fact that the speech is limited to a very narrow 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 (Kameyama & Arima, 1994), 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.

2.3.3.1 Summarization of Newscasts

One of the domains of speech summarization that has received the most attention and has perhaps 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. (1999), a user interface tool is provided for browsing and information retrieval of spoken audio - in this case, using TREC-7 SDR data (Voorhees & Harman, 1999). 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.

Another example of adding structure to speech data is in the work of Barzilay et al. (2000). The authors focus on classifying speaker roles in radio broadcasts, automatically discerning between anchors, journalists and program guests using lexical and

durational cues. This speaker role identification can be valuable for quickly indexing a large amount of broadcast data and especially for finding the transitions between stories.

In Valenza et al. (1999), summarization of the American Broadcast News corpus was 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 Chapter 4). The units of extraction are n-grams, utterances and keywords, which in the case of n-grams and utterances are scored according to the normalized sums of their constituent words. 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 (Murray et al., 2005a). Valenza et. al also provide a simple but intuitive interface for browsing the recognizer output.

In work by Hori and Furui (2000) 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 reported that 86% of the important words in the test set are extracted.

Kolluru et al. (2005) used a series of multi-layer perceptrons to summarize newscasts, by removing ASR errors according to recognizer confidence scores and then selecting units at increasing levels of granularity, based on term-weighting and Named Entity features. They found that their summarizer performed very well according to a question-answering evaluation and ROUGE analysis, but slightly less well on subjective fluency criteria.

More recently in the broadcast news domain, Maskey and Hirschberg (2005) found that the best summarization results in this domain 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. The highest F-measure reported was 0.544. ROUGE recall scores were also reported, with ROUGE-2 scores as high as 0.80 and ROUGE-SU scores as high as 0.75. Acoustic/prosodic and structural features alone yield ROUGE scores in the range of 0.68-0.76. Work by Ohtake et al. (2003) 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.

Christensen et al. (2008) have developed a system for skimming broadcast news transcripts, consisting of three steps of automatic speech recognition, story and utterance segmentation, and determination of the most informative utterances, which are then highlighted in the transcript. Salience is determined by features of position, length, *tf.idf* score and cosine similarity of utterance and story term-vectors. They evaluated their system both intrinsically with recall, precision and f-score, and extrinsically via a question-answering task. Two relevant findings are that ASR did not seriously affect the determination of salience, but that errors in story segmentation had a detrimental impact on downstream processes.

2.3.3.2 Summarization of Meetings

In the domain of meetings, Waibel et al. (1998) implemented a modified version of MMR 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 (2002) 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. This work also focused on detecting disfluencies such as filled pauses, false starts and repairs in order to increase summary readability and informativeness. Summarization Accuracy scores were reported, ranging from 0.506 to 0.614 in the various dialogue corpora.

On the ICSI corpus, Galley (2006) 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 features* 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. Galley reported ROUGE-2 scores in the range of 0.42-0.44 and Pyramid scores in the range of 0.504-0.554.

Simpson and Gotoh (2005), 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-independent 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. (2007) 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. Zhu and Penn (Zhu & Penn, 2006) showed how disfluencies can be exploited for summarization purposes and found that non-lexicalized filled-pauses were particularly effective for summarizing SWITCHBOARD speech. ROUGE-1 scores range between 0.502 for 30% utterance-based compression to 0.628 for 10% compression.

In our own work on the ICSI corpus, Murray et al. (2005a, 2005b) compared text summarization approaches with feature-based approaches incorporating prosodic features, with human judges favoring the feature-based approaches. In subsequent work (Murray et al., 2006), 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 judgements on the ICSI test data.

2.3.3.3 Summarization of Lectures

Hori et al. (2003) 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. Summarization accuracy results (word accuracy between an automatic summary and the most similar string from the referent summary word network) were reported, with scores in the range of 25-40 for a 50% summarization ratio and 35-56

for the 70% summarization ratio.

Also in the lectures domain, Fujii et al. (2007) 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. (2007) 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.

2.3.3.4 Voicemail Summarization

The SCANMail system (Hirschberg et al., 2001) 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. (2001) and Jansche and Abney (2002) also described techniques for extracting phone numbers from voicemails.

Koumpis and Renals (2005) 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.

2.3.4 Summarization Evaluation

Summarization evaluation techniques can generally be classified as *intrinsic* or *extrinsic* (Jones & Galliers, 1995). Intrinsic metrics evaluate the actual information content of a summary, usually by comparing it either with gold-standard human summaries or with the full document source. Extrinsic metrics, on the other hand, evaluate the usefulness of the summary in performing a real-world task. Most summarization work to date has relied much more heavily on intrinsic measures than extrinsic measures, for the primary reason that such evaluations are more easily replicable and subsequently more useful for development purposes. Here we consider the most widely used intrinsic summarization evaluation techniques to date, and we save a discussion of extrinsic

approaches until Chapter 6 (page 93), where we place our own extrinsic evaluation in the context of previous evaluations.

A definitive overview of summarization evaluation techniques is difficult if not impossible, as the summarization community has never agreed on an intrinsic evaluation framework and researchers have tended to rely on their own in-house metrics. In recent years, however, a suite of evaluation metrics under the name ROUGE has become increasingly popular (Lin & Hovy, 2003). ROUGE in turn is a variation of BLEU (Papineni et al., 2001), a machine translation evaluation tool. BLEU is based on comparing n-gram overlap between machine translations and multiple gold-standard human translations and is precision-based. ROUGE was developed essentially as a recall-based version of BLEU, though the most recent versions of ROUGE calculate precision, recall and f-score. There are several metrics within the ROUGE suite, but the most widely used are ROUGE-2 and ROUGE-SU4, the former of which calculates bigram overlap and the latter of which calculates skip bigram overlap with up to four intervening terms. Lin (2004) provided evidence that these metrics correlate well with human evaluations for several years' worth of DUC data. Subsequent research has yielded mixed results concerning ROUGE correlations with human evaluations (Dorr et al., 2004; Murray et al., 2005b; Dorr et al., 2005; Murray et al., 2006), but ROUGE has become an official metric of the Document Understanding Conference and is increasingly relied upon by researchers, allowing them to directly compare summarization results on given datasets.

The creators of ROUGE have also developed the Basic Elements evaluation suite (Hovy et al., 2006), which attempts to remedy the drawbacks of relying on n-gram units or sentence units for comparing machine summaries to reference summaries. Instead of relying on n-grams like ROUGE does, this evaluation framework uses units called Basic Elements, which are defined in the most simple case as either heads of major syntactic constituents (a single item) or relations between heads and dependents (a triple of head, modifier, and relation). The advantage of Basic Elements is that it features a deeper semantic analysis than simple n-gram evaluation, but the disadvantage is that it relies on parsing and pruning, which can be very problematic for disfluent speech data. Like ROUGE, Basic Elements is not a single evaluation metric. Rather it consists of numerous modules relating to three evaluation steps of *breaking*, *matching* and *scoring*, which correlate to locating the basic elements, matching similar basic elements, and scoring the summaries, respectively.

The Pyramid method (Nenkova & Passonneau, 2004) uses variable-length sub-

sentential units for comparing machine summaries to human model summaries. These *semantic content units* (SCUs) are derived by having human annotators analyze multiple model summaries for units of meaning, with each SCU being associated with a weight relating to how many model summaries it occurs in. These varying weights lend the model the pyramid structure, with a small number of SCUs occurring in many model summaries and most SCUs appearing in only a few model summaries. Machine summaries are then annotated for SCUs as well and can be scored based on the sum of SCU weights compared with the sum of SCU weights for an optimal summary. Using the SCU annotation, one can calculate both recall-based and precision-based summary scores. The advantage of the Pyramid method is that it uses content units of variable length and weights them by important according to occurrence in model summaries, but the disadvantage is that the scheme requires a great deal of human annotation. Pyramids were used as part of the DUC 2005 evaluation, with numerous institutions taking part in the peer annotation step, and while the submitted peer annotations required a substantial amount of corrections, Nenkova et al. (2007) reported acceptable levels for inter-annotator agreement. Galley (2006) introduced a matching constraint for the Pyramid method, namely that when comparing machine extracts to model extracts, SCUs are only considered to match if they originate from the same sentence in the transcript. This was done to account for the fact that sentences might be superficially similar in each having a particular SCU but nevertheless have much different overall meanings.

Work on *factoid*-based evaluation by Teufel and van Halteren (2004) is similar to the Pyramid method, except that factoids are atomic units whereas SCUs are of variable length and can be quite long. Additionally, factoid weights can be determined by features beyond frequency, such as document position. The sentence "Police have arrested a white Dutch man" is represented by the following factoids provided by Teufel and Van Halteren:

- A suspect was arrested.
- The police did the arresting.
- The suspect is white.
- The suspect is Dutch.
- The suspect is male.

The inter-annotator agreement for factoid annotation on news-wire data was quite high according to the kappa value, at around 0.86.

There has also been knowledge transfer between the question-answer (QA) and summarization domains in recent years. In the TREC QA track (Voorhees, 2004), non-factoid questions, i.e. questions that require lengthier responses, are evaluated using *information nuggets*, which are automatic information units that are considered by an assessor to be relevant to the information need. Interestingly, the actual system responses are used by the assessor in identifying the information nuggets. These nuggets are identified as vital or non-vital, and systems are scored with nugget precision and nugget recall, deriving an overall f-score. Nenkova et al. (2007) discussed how nuggets and pyramids might be used together, and additional ideas for knowledge transfer between these domains was provided by Lin and Demner-Fushman (2005, 2006).

The weighted precision metric (Murray et al., 2006) can be seen as being analogous to the Pyramid method, but with dialogue acts as the SCUs. This evaluation metric relies on human gold-standard abstracts, multiple human extracts, and links between the abstracts and extracts. The annotations and the evaluation scheme are described in detail in Chapter 3 (page 25), with the scheme extended from the original weighted precision to weighted precision/recall/f-score. The advantage of the scheme is that once the model annotations have been completed, new machine summaries can easily and quickly be evaluated, but the disadvantage is that it is limited to evaluating extractive summaries and works only at the dialogue act level.

Radev and Tam (2003) proposed a somewhat similar evaluation method to weighted precision for extractive summarization, *relative utility*. Human annotators are asked to rate each document sentence on a scale of 1 to 10, with 10 being the maximum score for meriting inclusion in the summary. Machine extracts are then evaluated according to how well the extracted sentences score according to the human judges, normalized by the maximum achievable score for the given summary length. Redundancy information is also explicitly marked, so that the inclusion of one sentence might penalize the presence of another. The advantage of this approach over simple precision and recall is that in the latter case, human sentence selection can be dependent on summary length (Jing et al., 1998; Mani, 2001b) and one sentence might be selected while a very similar sentence is not, whereas with relative utility all sentences are scored for extract-worthiness and the metric can be easily applied to summaries of various lengths. The disadvantage is basically the same as with weighted precision, that the method is only applicable to extractive summaries and does not operate at a level of fine granularity.

Zechner and Waibel (2000) introduced an evaluation metric specifically for speech summarization, *summarization accuracy*. The general intuition is that an evaluation method for such summaries should take into account the relevance of the units extracted as well as the recognition errors for the words which comprise the extracted units. Annotators are given a topic-segmented transcript and told to select the most relevant phrases in each topic. For summaries of recognizer output, the words of the ASR transcripts are aligned with the words of the manual transcripts. Each word has a relevance score equal to the average number of times it appears in the annotators' most relevant phrases. Given two candidate sentences, sentence 1 might be superior to sentence 2 when summarizing manual transcripts if it contains more relevant words, but if sentence 1 has a higher WER than sentence 2 it may be a worse candidate for inclusion in a summary of the ASR transcript. Summaries with high relevance and low WER will thereby rate more highly.

The challenge with evaluating summaries intrinsically is that there is not normally a single best summary for a given source document. Given the same input, human judges will often exhibit low agreement in the units they select (Mani et al., 1999; Mani, 2001b). In early work on automatic text summarization, Rath et al. (1961) showed that even a single judge who summarizes a document once and then summarizes it again several weeks later will often create two very different summaries (In that specific case, judges could only remember which sentences they had previously selected 42.5% of the time). With many annotation tasks, such as dialogue act labeling for example, one can expect high inter-annotator agreement, but summarization annotation is clearly a more difficult task. As Mani et al. (1999) pointed out, there are similar problems regarding the evaluation of other NLP technologies that may have more than one acceptable output, such as natural language generation and machine translation. The metrics described above have various ways of addressing this challenge, relying generally on multiple references. With ROUGE, n-gram overlap between a machine summary and multiple human references is calculated, and it is assumed that a good machine summary will contain certain elements of each reference. With pyramids, the SCUs are weighted based on how many summaries they occur in, and with weighted f-score, we rely on multiple annotators' links between extracts and abstracts. Teufel and van Halteren (2004) and Nenkova et al. (2007) discussed the issue of how many references are needed to create reliable scores, but the crucial point is that there is no such thing as a single best summary and multiple gold-standard reference summaries are desirable. As Galley (2006) observed, the challenge is not low inter-annotator agree-

ment itself but in using evaluation metrics that account for the diversity in reference summaries.

These are only a few of the evaluation metrics used in recent years, and each has advantages and disadvantages. What metrics like ROUGE, weighted precision, relative utility and summarization accuracy have in common is that there is an initial stage of manually creating model summaries, and subsequently new machine summaries can be quickly and automatically evaluated. In contrast, Pyramids, nuggets and factoids require additional manual annotation of machine summaries. On the other hand, these latter evaluation schemes operate at a more meaningful level of granularity compared to using n-grams or entire sentences. What all these schemes have in common is replicability, being able to reproduce the results once the relevant annotations have been done, which is not feasible when simply enlisting human judges to conduct subjective evaluations of summary informativeness or quality. Such human evaluations are very useful for periodic large-scale evaluation of summarization systems, however, and crucial for ensuring that automatic or semi-automatic metrics correlate with human judgements or real-world utility.

2.4 Further References

For further overviews of text summarization research and directions, see Mani (2001a), Jones (1999) and Endres-Niggemeyer (1998).

Chapter 3

Meeting Corpora and Experimental Design

The summarization experiments described herein are carried out on spontaneous multi-party spoken dialogues, or meeting speech. This is a particularly interesting speech domain because of the naturalness of the speech and the challenges presented by disfluent, overlapping dialogues. Domains such as broadcast news and lectures are popular among the summarization and information extraction communities, but represent a middle ground between text and speech data, as there is often a prepared and read aspect to the speech. Purely spontaneous speech can be dramatically different in terms of fluency, prosody and information density when compared with these other speech domains. This presents many challenges for automated analysis, from automatic speech recognition to automatic extraction of informative dialogue acts, the focus of this work. By working in this domain, we hope to discover the correlates of informativeness for speech in unscripted, natural settings.

We use two corpora for our experiments, the AMI and ICSI meeting corpora, described in detail below.

3.1 AMI Corpus

The AMI corpus consists of $\sim$100 hours of recorded and annotated meetings, divided into *scenario* and *non-scenario* meetings. In the scenario meetings, four participants take part in each meeting and play roles within a fictional company. The scenario given to them is that they are part of a company called Real Reactions, which designs remote controls. Their assignment is to design and market a new remote control, and the

members play the roles of project manager (the meeting leader), industrial designer, user-interface designer, and marketing expert. Through a series of four meetings, the team must bring the product from inception to market.

The first meeting of each series is the kick-off meeting, where participants introduce themselves and become acquainted with the task. The second meeting is the functional design meeting, in which the team discusses the user requirements and determines the functionality and working design of the remote. The third meeting is the conceptual design of the remote, wherein the team determines the conceptual specification, the user interface, and the materials to be used. In the fourth and final meeting, the team determines the detailed design and evaluate their result.

The participants are given real-time information from the company during the meetings, such as information about user preferences and design studies, as well as updates about the time remaining in each meeting. While the scenario given to them is artificial, the speech and the actions are completely spontaneous and natural. There are 138 meetings of this type in total. The length of an individual meeting ranges from $\sim$15 to 45 minutes, depending on which meeting in the series it is and how quickly the group is working.

The non-scenario meetings are meetings that occur regularly and would have been held regardless of the AMI data collection, and so the meetings feature a variety of topics discussed and a variable number of participants. For the experiments described in this thesis, we use only the scenario meetings from the AMI corpus.

The meetings were recorded at three locations: Edinburgh, TNO, and IDIAP. The participants consist of both native and non-native English speakers, and many of them are students.

The AMI corpus is freely available[1] and contains numerous annotations for a variety of multi-modal phenomena.

3.2 ICSI Corpus

The second corpus used herein is the ICSI meeting corpus (Janin et al., 2003), a corpus of 75 natural, i.e. non-scenario, meetings, approximately one hour each in length. As with the AMI non-scenario set, these are meetings that would have been held anyway and feature a variable number of participants. Because many of the meetings in the corpus are gatherings of ICSI researchers themselves, the topics tend to be specialized

[1] http://corpus.amiproject.org/

and technical, e.g. discussions of speech and language technology. The average length of an ICSI meeting is greater than the average AMI non-scenario meeting.

Like the AMI corpus, the ICSI corpus meetings feature both native and non-native English speakers. All meetings in the corpus were recorded at ICSI in Berkeley, California. Unlike the AMI scenario meetings and similar to the AMI non-scenario meetings, there are varying numbers of participants across meetings in the ICSI corpus, with an average of six but sometimes as many as ten per meeting.

Unlike the AMI corpus, which is multi-modal and contains a variety of information such as slides, whiteboard events and participant notes, the ICSI corpus consists entirely of speech and relevant annotations.

3.3 Human Annotation

This section gives an overview of the sets of manual annotation that are used throughout the experiments described in this thesis.

3.3.1 Dialogue Act Annotation

As described in the introduction, we are engaged here in the task of *extractive* summarization, wherein we classify certain segments from the source document as summary-worthy and reject the others, concatenating the chosen units into a single compressed document. Whereas the unit of extraction for text summarization might be a sentence, the unit of extraction for this spontaneous speech data is the *dialogue act*. In these meetings, as with other spontaneous speech corpora, people tend not to speak in complete and grammatical sentences, and so we instead segment the speech stream according to speaker intentions. Each dialogue act segment roughly corresponds to a single speaker intention. A dialogue act can contain more than one sentence-type unit or less than a whole sentence-type unit, since the segmentation is based primarily on intention rather than grammatical considerations. Annotators also label each dialogue act segment with a type, such as "back-channel," "inform," and "suggest," but for these experiments we use only the segmental information and disregard the dialogue act type.

Though we primarily use hand-segmented dialogue acts as the summarization units, in Chapter 8 (page 165) we explore the impact of using a much simpler pause-based spurt segmentation in lieu of dialogue act segmentation and survey the impact of this

simplified segmentation on the summarization task. Also on the AMI and AMIDA projects, Dielmann and Renals (2007) have researched automatic segmentation and labelling of dialogue acts for the AMI corpus.

3.3.2 Summarization Annotation

For both the AMI and ICSI corpora, annotators were asked to write abstractive summaries of each meeting and to extract the meeting dialogue acts that best convey or support the information in the abstractive summary.

Annotators used a graphical user interface (GUI) to browse each individual meeting, allowing them to view previous human annotations comprised of an orthographic transcription synchronized to the meeting audio, and topic segmentation. Some of these summarization annotators had previously taken part in the topic segmentation annotation while others were unfamiliar with the data. The annotators were first asked to build a textual summary of the meeting aimed at an interested third-party, using four headings for the summary. For the ICSI meetings, the four headings are:

- general abstract: "why are they meeting and what do they talk about?";
- decisions made by the group;
- progress and achievements;
- problems described

For the AMI meetings, the summary sections were slightly different:

- general abstract;
- decisions;
- actions;
- problems;

The maximum length for each summary section is 200 words, and while it was mandatory that each general abstract section contained text, it was permitted that for some meetings the other three sections could be null; for example, some meetings might not involve any decisions being made. Annotators who were unfamiliar with the

data were encouraged to listen to a meeting in its entirety before beginning to compose the summary.

After authoring the abstractive summary, annotators were then asked to create an extractive summary, using a second GUI. With this GUI they were able to view their textual summary and the orthographic transcription, with the topic segments removed and with one dialogue act per line based on the pre-existing MRDA coding (Shriberg et al., 2004). They viewed only the dialogue act segments without the dialogue act type labels. They were told to extract the dialogue acts that together could best convey the information in the abstractive summary and could be used to support the correctness of the abstract. They were not given any specific instructions about the number or percentage of dialogue acts to extract, nor any instructions about extracting redundant dialogue acts. They were then required to do a second pass annotation, wherein for each extracted dialogue act they chose the abstract sentences supported by that dialogue act. The result is a many-to-many mapping between abstract sentences and extracted dialogue acts, i.e. an abstract sentence can be linked to more than one dialogue act and vice-verse. Although the expectation was that each abstract sentence would be linked to at least one extracted dialogue act and each extracted dialogue act linked to at least one abstract sentence, annotators were permitted to leave abstract sentences and dialogue acts standing alone in some circumstances. However, for training our statistical models in Chapters 5, 6 and 7, only dialogue acts that are linked to abstract sentences are considered to be members of the positive class. This is done to maximize the likelihood that a data point labelled as "extractive" is truly an informative example for training purposes; on average, fewer than 10% of the dialogue acts extracted by an annotator remain unlinked. Note that in this research the number of dialogue act links is used only for evaluation purposes. For training our binary classifiers, we simply consider a dialogue act to be a positive example if it is linked to a given human summary, and a negative example otherwise. Future work could look at incorporating the link counts in a linear regression model.

For the test set meetings in each corpus, we also had multiple annotators write abstracts of the meetings so that we have multiple gold-standard summaries for evaluation purposes. For each AMI test set meeting, there are two human-authored abstract summaries. For each ICSI test set meeting, there are either four or five human-authored abstract summaries.

3.3.2.1 Annotator Agreement

To gauge inter-annotator agreement on the extractive coding, we can utilize the kappa statistic (Carletta, 1996). The kappa statistic is a way of evaluating how closely two annotators agree with each other on an annotation task. The statistic is derived by calculating

$$\kappa = (\text{Observed Agreement} - \text{Chance Agreement}) / (1 - \text{Chance Agreement})$$

For each meeting in the corpus, the kappa value for each annotator pair is calculated and these values are averaged to derive a single kappa value for that meeting. These averages are then summed and averaged over the corpus to derive an average kappa statistic for the corpus.

For the ICSI test set, the average kappa value is 0.35. For the AMI test set, the average kappa value is 0.48. Both scores are somewhat low, but as discussed in Chapter 2 (page 19), it is not unusual to have low annotator agreement for summarization annotation as there is normally no single best summary for a given document. We also find that the AMI test set agreement is considerably higher than the ICSI test set agreement, reflecting the difficulty in annotating the less structured ICSI meetings.

Whereas the ICSI corpus only has multiple extractive codings for the test set, we have multiple extractive codings for the entirety of the AMI scenario meetings. The annotator agreement for the entire AMI corpus is 0.45, slightly lower than for the test set alone.

3.4 Automatic Speech Recognition

For the experiments described in this thesis, we make extensive use of automatic speech recognition (ASR) for the two meeting corpora. This ASR output was provided by the AMI-ASR team (Hain et al., 2007). The AMI automatic transcription system uses the standard framework of hidden Markov model (HMM) acoustic modelling and n-gram language models, in this case tri-grams. To achieve fair recognition output, the corpus is divided into five parts, employing a leave-one-out procedure of training the language and acoustic models on four portions of the data and testing on the fifth, rotating to obtain recognition results for the entire corpus (for the ICSI data this is four parts rather than five).

The microphones used for this speech recognition output are individual headset microphones. The AMI ASR system features components for crosstalk suppression

and automatic segmentation using a multi-layer perceptron (MLP).

The WER for the ICSI corpus is 29.5% and the WER for the AMI corpus is 38.9%. There are multiple versions of ASR available in AMI corpus and the version used here is labelled *ASR_AS_CTM_v2.0_feb07* in the corpus, and is distinct from the other available versions of the ASR in that it features automatic segmentation. This system also incorporates Vocal Tract Length Normalization (VTLN) and Maximum Likelihood Linear Regression (MLLR) adaptation.

3.5 Experimental Overview

This section provides a description of the experimental setup for this set of summarization experiments.

3.5.1 Training, Development and Test Sets

For the experiments on the AMI data, the corpus was divided into three portions: training, development, and test data. All the AMI meetings used were taken from the scenario portion of the corpus. The training data consists of 92 meetings, the development set contains 24 meetings, and the test set is comprised of 20 meetings, or five meeting series. The test set consists of meetings recorded at multiple AMI facilities: eight recorded in Edinburgh, four recorded at IDIAP, and eight recorded at TNO.

The ICSI training set consists of 69 meetings and the test set is comprised of 6 meetings.

3.5.2 Extractive Classifiers

For the supervised classification experiments described in Chapters 5 and 7, the classifier used is the *liblinear* logistic regression classifier[2]. This classifier type is useful and efficient for binary classification tasks and for training on large datasets. The logistic regression probability model is given by

$$P(y = \pm 1|x) = \frac{1}{1 + exp(-y(a + bx))}$$

where x represents the data, a and b are weights estimated by maximum likelihood, and y is the class label. The *liblinear* toolkit incorporates simple feature subset

[2]http://www.csie.ntu.edu.tw/ cjlin/liblinear/

selection based on calculating the f statistic for each feature and performing cross-validation with subsets of various sizes, comparing the resultant balanced accuracy scores. The f statistic for each feature is calculated in *liblinear* by the formula

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

where n_+ and n_- are the number of positive instances and negative instances, respectively, $\bar{x}_i, \bar{x}_i^{(+)}$, and $\bar{x}_i^{(-)}$ are the means of the ith feature for the whole, positive and negative data instances, respectively, $x_{k,i}^{(+)}$ is the ith feature of the kth positive instance, and $x_{k,i}^{(-)}$ is the ith feature of the kth negative instance (Chen & Lin, 2006).

In preliminary work we used an SVM classifier with an RBF kernel, trained on the same data and using the same feature sets, and this was the classifier used to create the summaries described in Chapter 6, the extrinsic evaluation discussion. However, there was not a noticeable performance difference between using SVMs and logistic regression classifiers, and the latter classifier is much faster to train, so we ultimately chose the logistic regression classifier for the bulk of our experiments in order to expedite development. Chapter 6 represents the only use of SVMs in this thesis.

3.5.3 Compression Level

For each summarization system presented in Chapters 4, 5 and 7, we create summaries with a length of 700 words each. This length is chosen so that the summaries could hypothetically satisfy two use cases: they are brief enough to be read by a time-constrained user, much as an abstractive summary might be quickly reviewed, but long enough to serve as indices into the most important points of the meeting records. This short summary length also necessitates a high level of precision since we extract relatively few dialogue acts. For the decision audit task described in Chapter 6, the summaries of the four relevant meetings are of a length approximating the length of the manual extracts for each of those meetings. This is done because the gold-standard human abstract condition in that experiment contains links to the human-extracted dialogue acts in the transcript, and we want extractive coverage to be comparable.

Further research is needed to determine the optimal compression levels for summarization of such data. Most systems to date extract units until reaching a limit defined by a preset percentage of sentences, preset percentage of words or a preset word-count, the latter being the method used herein. All methods seem to have their disadvantages. For example, when humans create extractive summaries of meetings, longer meetings

do yield longer summaries in general, but the word-count percentage actually decreases as the meetings get longer. That is, the word-count for the summary of a short meeting will tend to represent a much higher percentage of the total meeting word-count than will the summary of a very long meeting.

In any case, for our most in-depth analyses of the systems, we extrapolate away from any specific compression ratio or posterior probability threshold by evaluating the receiver operator characteristic (ROC) curves and the areas under the curve (AUROC) for the various classifiers including feature-subset classifiers. This allows us to evaluate how well the summarization systems discern true positives and false positives, regardless of summary length.

3.5.4 Evaluation

This section describes the summarization evaluation schemes used throughout these experiments. We first introduce and provide details for two intrinsic evaluation metrics, and subsequently motivate an extrinsic evaluation that is described in detail in Chapter 6.

3.5.4.1 Weighted Precision, Recall and F-Score

In previous work (Murray et al., 2006) we introduced the weighted precision evaluation. Here we extend that analysis to weighted precision, recall and f-score. While evaluation metrics such as ROUGE, which work at the word or n-gram level, are primarily or originally *recall* measures, we have previously assumed that when performing a weighted evaluation at the dialogue act level, *precision* is more informative, particularly when the summaries are moderately or severely short. However, we derive the f-score here for completeness and present that as the central evaluation metric of interest.

To calculate weighted precision, we count the number of times that each extractive summary dialogue act was linked by each annotator, averaging these scores to get a single dialogue act score, then averaging all of the dialogue acts scores in the summary to get the weighted precision score for the entire summary. To calculate weighted recall, the total number of links in our extractive summary is divided by the total number of links to the abstract as a whole. A difference between weighted precision and weighted recall is that weighted recall has a maximum score of 1, in the case that all linked dialogue acts are included in the extractive summary, whereas there is no the-

oretical maximum for weighted precision since annotators were able to link a given dialogue act as many times as they saw fit.

More formally, both weighted precision and recall share the same numerator

$$num = \sum_{i=1}^{M} \sum_{j=1}^{N} L(s_i, a_j)$$

where $L(s_i, a_j)$ is the number of links for a dialogue act s_i in the machine extractive summary according to annotator a_i, M is the number of dialogue acts in the machine summary, and N is the number of annotators.

Weighted precision is equal to

$$precision = \frac{num}{N \cdot M}$$

Weighted recall is given by

$$recall = \frac{num}{\sum_{i=1}^{O} \sum_{j=1}^{N} L(s_i, a_j)}$$

where O is the total number of dialogue acts in the meeting, N is the number of annotators, and the denominator represents the total number of links made between dialogue acts and abstract sentences by all annotators.

The f-score is calculated as the harmonic mean of precision and recall:

$$(2 * precison * recall)/(precision + recall)$$

In general, these weighted metrics are based on the assumption that dialogue acts that are linked multiple times by multiple annotators are more informative and should be weighted more highly when included in a summary.

3.5.4.2 ROUGE

In Chapter 2 (page 19) we describe the ROUGE intrinsic evaluation metric. In previous experiments (Murray et al., 2005b, 2006) we found that the ROUGE metrics did not correlate well with subjective human judgements of summaries of the ICSI meeting test set, and so ROUGE is not used as the primary evaluation in these further experiments since we can rely on the human extractive gold-standards using the weighted f-score scheme described above. However, the disadvantage of weighted precision, recall and f-score and a potential advantage of ROUGE is that the former is restricted to the evaluation of extractive summaries whereas ROUGE can be used to compare any type

of automatic summary with human reference summaries, since it works at the n-gram level rather than the dialogue act level.

Though the research in this thesis mostly relies on weighted precision/recall/f-score for evaluation purposes, we do utilize ROUGE as an evaluation metric in Chapters 7 and 8 in addition to weighted f-score. In those instances, we rely on the ROUGE-2 and ROUGE-SU4 metrics, which calculate bigram overlap and skip bigram overlap with up to four intervening words, respectively.

3.5.4.3 Extrinsic Evaluation

In Chapter 2 (page 19) we describe the difference between extrinsic and intrinsic approaches to summarization evaluation. This thesis argues that truly robust summarization evaluation will incorporate extrinsic measures in addition to intrinsic measures. While intrinsic evaluation metrics are indispensable for development purposes and can be easily replicated, they ideally need to be chosen based on whether or not they are good predictors for extrinsic usefulness, e.g. whether they correlate to a measure of real-world usefulness. Evaluating according to human gold-standard annotations is sensible, but ultimately all summarization work is done for the purpose of facilitating some task and should be evaluated in that context. As Sparck-Jones has said, "it is impossible to evaluate summaries properly without knowing what they are for" (Jones, 1999). Ideally, even evaluation measures that compare a summary with a full source document or a model summary would do so with regards to use constraints. Here we directly evaluate the utility of the summaries for a particular use case by measuring their impact on satisfying a typical information need relating to that use case.

Specifically, our incorporation of extrinsic measures in this thesis is related to our domain of speech summarization and to the predicted use cases of the meeting summaries generated. The summaries are meant to be used in the context of a meeting browser, aiding a time-restricted user who needs to quickly review meeting content for use cases such as preparing for a subsequent meeting or plumbing corporate memory. In these cases, it is not sufficient merely to know that our automatically generated summaries are to some degree similar to manually drafted summaries, as the documents are not intended to be stand-alone documents. Rather, they are included in a meeting browser as a navigational tool. For example, a user of the meeting browser can first read the extractive summary in its entirety and then navigate the entire transcript and audio/video record by clicking on summary dialogue acts as an index into the record. It is crucial, therefore, to know just how good extractive summaries are as navigational

tools for such purposes. Figure 3.1 illustrates the relationship between an extractive summary and the overall meeting record. Ultimately summarization may be merely one component of a multi-media meeting browser, but here we want to isolate the impact of summarization compared with other possible components or configurations. We are interested in how well we meet the needs of a particular use-case when each individual information component is featured.

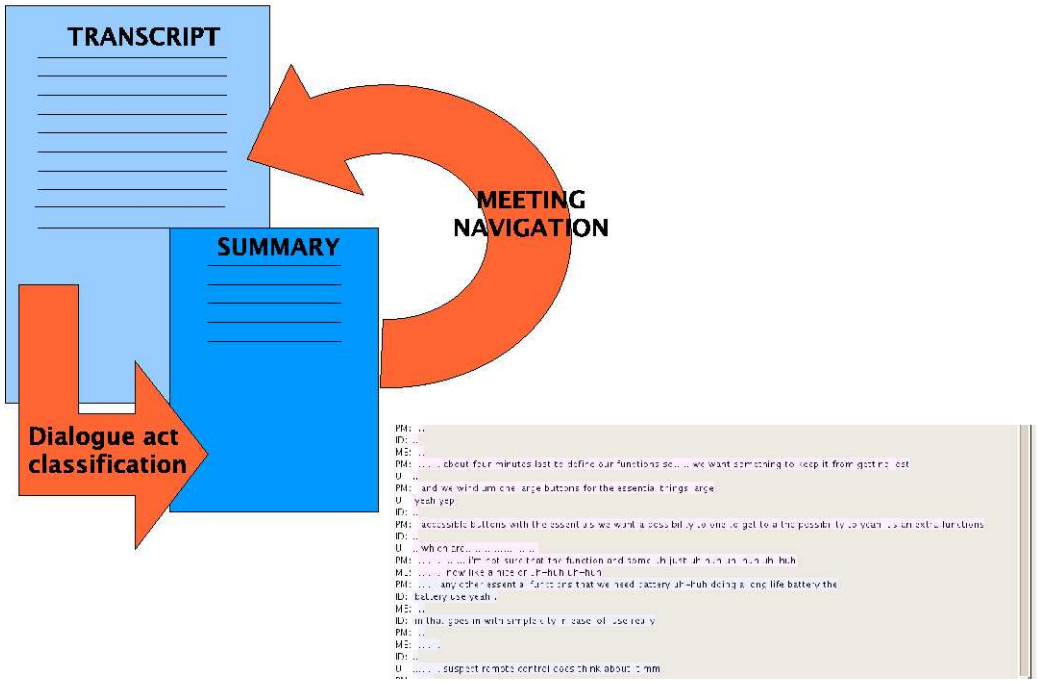


Figure 3.1: Summaries as Navigation Aids

Another way of motivating such an extrinsic evaluation for our purposes, and placing it in relation to intrinsic evaluations, is that our intrinsic evaluations tell us how multiple extraction techniques compare to one another, while the extrinsic evaluation tells us how useful the extrinsic paradigm is as a whole. It is of little use to say that our extractive summaries are very good compared with gold standard human extracts if they are not useful as navigational tools in a meeting browser. It may be the case that we can successfully locate informative dialogue acts in a meeting, but that users find it very difficult to read sentences that have been removed from their original contexts and concatenated together, or that the ASR errors make reading comprehension a great challenge. We therefore emphasize the importance of the real-world application of our data. Of course, the decision audit task could be set up in such a way that it compares multiple approaches to extractive summarization with one another, but here we more

generally evaluate how useful extractive summarization is for the particular use-case.

Chapter 6 (page 93) describes our large-scale extrinsic evaluation for automatic summarization. The particular form of the evaluation is a *decision audit* task, wherein a user must review several archived meetings in order to satisfy a complex information need. The task is described in detail and placed in the context of previous extrinsic evaluations.

3.6 Conclusion

This chapter has served to provide a description of the corpora used and the general experimental overview. The individual chapters supply further experimental details where appropriate.

Chapter 4

Keywords and Cuewords

This chapter examines the use of keywords and cuewords in speech summarization. In the first section we survey established term weighting methods, discuss our implementations of several state-of-the-art term weighting techniques in a simple summarization system, and introduce two novel term weighting methods for spontaneous spoken dialogues. The second section examines the usefulness of cuewords for speech summarization. The difference in keywords and cuewords is that the latter are somehow specific to a given document and to the topics within the document, and are therefore useful for distinguishing documents, whereas cuewords are more generally indicative of informative areas of a document and are not necessarily specific to a document at hand. The latter are more common words that could signal informativeness across a variety of documents and topics.

4.1 Term Weighting

In this section we explore a variety of term weighting techniques for spontaneous speech data in the meetings domain. In term weighting, we assign scores to each word in a document so that the most informative end up with the highest scores and less informative words and function words have scores at or near zero. Several such weighting schemes are discussed below, including two novel techniques intended specifically for multi-party spoken dialogues. Choosing and implementing a term weighting method is often the first step in building an automatic summarization system. Though the unit of extraction may be the sentence or the dialogue act, those units are normally weighted by the importance of their constituent words. Popular text summarization techniques such as Maximal Marginal Relevance (MMR) and Latent Semantic Analy-

sis (LSA) begin by representing sentences as vectors of term weights. There is a wide variety of term weighting schemes available, from simple binary weights of word presence/absence to more complex weighting schemes such as *tf.idf* and *tf.ridf*. Several of these are described in the following section.

A central question of this section is whether term-weighting techniques developed for information retrieval (IR) and summarization tasks on text are well-suited for our domain of multi-party spontaneous spoken dialogues, or whether the patterns of word usage in such dialogues can be exploited in order to yield superior term-weighting for our task. To this end, we devise and implement two novel term-weighting approaches for multi-party speech, based on features such as differing word frequencies among speakers in a meeting and the relationship between keywords and meeting structure. These metrics are compared with 4 popular term-weighting schemes - *idf*, *tf.idf*, *ridf* and *Gain* - and the metrics are evaluated via an extractive summarization task on both AMI and ICSI corpora.

4.1.1 Previous Term Weighting Work

Term weighting methods form an essential part of most IR systems. Terms that characterize a given document well and discriminate the document from the remainder of the document collection should be weighted highly (Salton & Buckley, 1988). The most popular term weighting schemes have therefore combined *collection frequency* metrics with *term frequency* metrics.

The most common method of calculating collection frequency is called the *inverse document frequency* (IDF) (Jones, 1972). The IDF for term t is given by

$$IDF(t) = \log D - \log D(t)$$

or equivalently,

$$IDF(t) = -\log(\frac{D(t)}{D})$$

where D is the total number of documents in the collection and $D(t)$ is the number of documents containing the term t. A term will therefore have a high IDF score if it only occurs in a few documents in the document collection.

For the *term frequency* component, the simplest method is a binary term weight: 0 if the term is not present and 1 if it is. More commonly, the number of term occurrences in the document is used. Thus the term frequency TF is given by

$$TF(t,d) = \frac{N(t)}{\sum_{k=1}^{T} N(k)}$$

where $N(t)$ is the number of times the term t occurs in the given document and $\sum_{k=1}^{T} N_k$ is the total word count for the document, thereby normalizing the term count by document length.

The classic method for combining these components is simply *tf.idf* (Rijsbergen, 1979; Salton & Buckley, 1988), wherein a term is scored highly if it occurs many times within a given document but rarely across the set of all documents, by multiplying TF and IDF. This term weighting scheme *tf.idf* increases our ability to discriminate between the documents in the collection. While there are variants to the TF and IDF components given above (Salton & Buckley, 1988), the motivating intuitions are the same. Another example of combining these three types of data (collection frequency, term frequency and document length) is given by Robertson and Jones (1994) and is called the Combined Weight. For a term t and document d, the Combined Weight is described as:

$$CW(t,d) = \frac{IDF(t) \cdot TF(t,d) \cdot (K+1)}{K \cdot ((1-b) + (b \cdot (NDL(d)))) + TF(t,d)}$$

where K is a tuning constant regulating the impact of term frequency, b is a tuning constant regulating the impact of document length, and NDL is the normalized document length.

When a query q is given to a document d, the document can be scored for query relevance using the so-called Okapi BM25 score (Robertson et al., 1998),

$$BM25(d,q) = \sum_{i=1}^{n} CW(q_i,d)$$

where $q_1 \ldots q_n$ are the query terms. This scoring method has been the most reliable text retrieval term-weighting scheme in the TREC conferences (Robertson et al., 1998; Jones et al., 2000; Craswell et al., 2005).

When relevance information is available, i.e. a subset of documents has been determined to be relevant to a user query, additional proven metrics are available for term relevance weighting and/or query expansion (Robertson & Jones, 1994). One example is the RSJ metric given in (Robertson & Jones, 1976):

$$RSJ(t,q) = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n-r}{N-n-R+r}\right)}$$

where R is the number of documents known to be relevant to the query q and r is the number of relevant documents containing term t. The following variation is sometimes used instead, partly to avoid infinite weights under certain conditions:

$$RW(t,q) = \log(\frac{((r+0.5)(N-n-R+r+0.5))}{((n-r+0.5)(R-r+0.5))}$$

In practice, however, there is little or no relevance information available when doing term weighting. Work by Croft and Harper (1979) has shown that IDF is an approximation of the RSJ relevance weighting scheme when complete relevance information is unavailable. Robertson (Robertson, 2004) further discusses the relationship between IDF and relevance weighting and places the IDF scheme on strong theoretical ground.

One extension of *idf* called *ridf* (Church & Gale, 1995) has proven effective for automatic summarization (Orasan et al., 2007) and named entity recognition (Rennie & Jaakkola, 2005). In *ridf*, the usual IDF component is substituted by the difference between the IDF of a term and its expected IDF according to the Poisson model. *ridf* can be calculated by the formula

$$expIDF(t) = -\log(1 - e^{(-f_t/D)})$$

$$ridf(t) = IDF(t) - expIDF(t)$$

where f_t is the frequency of the word across all documents D in the document collection. Church and Gale (1995) give the example of the words "boycott" and "somewhat", which have similar IDF scores for a corpus of Associated Press articles. Out of more than 85,000 documents, "boycott" occurs in 676 and "somewhat" occurs in 979, resulting in IDF scores of 7.0 and 6.4. Given the frequency of "somewhat", it would be expected to occur in 1007 documents according to chance as modelled by the Poisson, a number only slightly higher than the actual number of documents. In contrast, "boycott" is expected to occur in 1003 documents given its frequency, a number much higher than the 676 documents it actually occurs in. The distribution of the two words among documents is much different, as keywords will tend to cluster into a smaller number of documents. This divergence between expectation according to the Poisson and the actual number of documents indexed by the term can be used to adjust the IDF score. The authors also show that the mid-frequency terms tend to have the largest divergence from expectation.

Papineni (2001) also provides an extension to IDF. The author argues that the IDF of a word is not synonymous with the *importance* of a word, but is rather an optimal

weight for document self-retrieval; they are ideal weights for measuring document similarity, but not necessarily indicative of term importance. Papineni proposes a term-weighting metric *Gain* which is meant to measure importance or information gain of the term in the document:

$$Gain(t) = \frac{D(t)}{D}(\frac{D(t)}{D} - 1 - \log\frac{D(t)}{D})$$

Very common and very rare words have low gain; this is in contrast with IDF, which will tend to give high scores to uncommon words. For example, if our document collection consists of 100 documents, a term that occurs in 2 documents has a Gain score of 0.059, a term that occurs in 20 documents has a score of 0.162, and a term that occurs in 80 documents has a score of 0.019. As mentioned above, *ridf* also favors medium-frequency words (Orasan et al., 2007). As Papineni (2001) points out, the effective performance of metrics such as *ridf* and *Gain* seems to corroborate Luhn's observation that medium-frequency words have the optimal "resolving power" (Salton & McGill, 1983).

Mori (2002) introduce a term weighting metric for automatic summarization called Information Gain Ratio (IGR). The underlying idea of IGR is that documents are clustered according to similarity, and further grouped into sub-clusters. If the information gain of a word increases after clusters are partitioned into sub-groups, then it can be said that the word contributes to that sub-cluster and should thus be rated highly.

Finally, Song et al. (2004) introduce a term weighting scheme for automatic summarization that is based on lexical chains. Building lexical chains in the manner of Barzilay and Elhadad (1997), they weight chains according to how many word relations are in the chain, and weight each word in a chain according to how connected it is in the chain. On DUC 2001 data, they reported outperforming *tf* and *tf.idf* weighting schemes.

4.1.2 Term Weighting for Multi-Party Spoken Dialogue

This section describes two approaches towards designing a term-weighting scheme specifically for spontaneous multi-party spoken dialogues.

4.1.2.1 The *su.idf* metric

A common theme of most of the term-weighting metrics described in the previous section is that the distribution of words across a collection of documents is key to determining an ideal weight for the words. In general, words that are unique to a given

document or cluster of documents should be weighted more highly than words that occur evenly throughout the entire document collection. For example, *tf.idf* scores words highly if they occur many times in the relevant document but rarely across a set of all documents. For multi-party spoken dialogue, we have another potential source of variation in lexical usage: the speakers themselves. We introduce a new term weighting score for multi-party spoken dialogues by also considering how term usage varies across speakers in a given meeting. The intuition is that keywords will not be used by all speakers with the same frequency. Whereas *tf.idf* compares a given meeting to a set of all meetings, we can also compare a given speaker to a set of other speakers in the meeting. For each of the four speakers in a meeting, we calculate a surprisal score for each word that speaker uttered, which is the negative log probability of the term occurring amongst the other three speakers. The surprisal score for each word t uttered by speaker s is

$$surp(s,t) = -\log\left(\frac{\sum_{s' \neq s} TF(t,s')}{\sum_{r \neq s} N(r)}\right)$$

where $TF(t,s')$ is the term frequency of word t for speaker s' and $N(r)$ is the total number of words spoken by each speaker r. For each term, we total its speaker surprisal scores and divide by the total number of speakers to find the overall surprisal score $totsurp(t)$. Thus the surprisal score for a word is given by

$$totsurp(t) = \frac{1}{S}\sum_{i=1}^{S} surp(s_i,t)$$

where S is the total number of speakers in the meeting. So if the word *budget* is uttered once by speaker A, twice by speaker B, none by speaker C and ten times by speaker D, and each speaker says 100 words total, the surprisal score of *budget* for speaker D is $-\log(3.0/300)$, or 6.64. These individual surprisal scores are then summed and averaged over each speaker. Table 4.1 gives an example for three terms, showing the normalized term frequency for each speaker and the overall surprisal score for the term. The term "kinetic" has a high score, as it is used by speaker 2 the most, less often by speakers 1 and 3, and barely at all by speaker 4. The term "standard" likewise scores highly, as it is used primarily by speakers 1 and 2 and much less often for the other two speakers. The word "charger" scores much lower; it is spoken only by speaker 2, and so while the speaker surprisal score for speaker 2 will be high, the remainder of the speaker surprisal scores will be 0 and the average will therefore be low (if speaker s does not utter word t, then $surp(s,t)$ is 0).

This surprisal score, the first component of the term-weighting metric, is then multiplied by $\frac{s(t)}{S}$, where $s(t)$ is the number of speakers who speak that word and S is the total number of speakers in the meeting. The third component of the metric is the inverse document frequency, or IDF. The equation for IDF is

$$IDF(t) = -\log(\frac{D_t}{D})$$

where D is the total number of documents and D_t is the number of documents containing the term t. Putting these three components together, our term weighting metric is

$$su.idf(t) = totsurp(t) \cdot \frac{s(t)}{S} \cdot \sqrt{IDF(t)}$$

One motivation for this novel term weighting scheme is that many important words in such meeting corpora are not necessarily rare across all documents, e.g. *cost*, *design* and *colour*. They are also not necessarily the most frequent content words in the meetings. They would therefore not score highly on either component of *tf.idf*. Though we retain inverse document frequency for our new metric, the square root of IDF is used to lower its overall influence within the metric, so that a term will not necessarily be weighted low if it is fairly common or weighted high simply because it is rare. Since we also run IDF as a metric of its own, we can determine its contribution to *su.idf*.

The hypothesis is that more informative words will be used with varying frequencies between the four meeting participants, whereas less informative words will be used fairly consistently by all. It is possible that lexical entrainment, the phenomenon where speakers subtly imitate each other's word choices, could be a confounding factor by making lexical distinctions across speakers less defined, but we hypothesize that there will still be interesting lexical differences between speakers. The component $\frac{s(t)}{S}$ is included for two reason. First, because individuals normally have idiosyncrasies in their speaking vocabularies, e.g. one meeting participant might use a type of filled pause not used by the others or otherwise frequently employ a word that is particular to their idiolect. And second, a word that is used by multiple speakers but with much different frequency should be more important than a word that is spoken by only one person.

There are several reasons for hypothesizing that use of informative words will vary between meeting participants. One is that meeting participants tend to have unique, specialized roles relevant to the discussion. In the AMI corpus, these roles are explicitly labelled, e.g. "marketing expert." With a given role comes a vocabulary associated with that role, e.g. "budget" and "cost" would be associated with a finance expert and

Term	surp(w)	TF(w,s1)	TF(w,s2)	TF(w,s3)	TF(w,s4)
"kinetic"	9.50047550237	0.00177304964539	0.00196335078534	0.00135869565217	0.000476871721507
"charger"	2.4767226489	0.0	0.00065445026178	0.0	0.0
"standard"	9.07984240343	0.00177304964539	0.00130890052356	0.00407608695652	0.000476871721507

Table 4.1: Overall Surprisal Score and Normalized Term Frequencies for Each Speaker for 3 Terms

"scroll" and "button" would be associated with an interface designer. Second, even when the roles are not so clearly defined, different participants have different areas of interest and different areas of expertise, and we expect that their vocabularies reflect these differences.

For ease of reference, we subsequently refer to this first speech-specific metric as *su.idf*.

4.1.2.2 The *twssd* metric

Subsequent to the original work on *su.idf* (Murray & Renals, 2007), it was determined that there was a simpler way of conceptualizing the intuition behind that term-weighting method, and a more straight-forward term-weighting method based on that conceptualization. The underlying question is if we draw a term from a meeting at random, how confidently can we predict the speaker of that term? Our hypothesis is that keywords will be more closely linked with a single speaker. For each term t in a meeting, if we have calculated the conditional probabilities of the term given each speaker, it is easy to calculate the probability of each speaker S given the term t, using Bayes' Theorem, estimating the probabilities using the counts from the data:

$$p(S|t) = \frac{p(t|S)p(S)}{p(t)}$$

We can then take the maximum of these speaker probabilities as our score Sc_1, representing our confidence that we can identify the speaker of the word at hand.

$$Sc_1(t) = \max_S p(S|t)$$

Also in our previous work on *su.idf*, we hypothesized that additional features for spontaneous spoken dialogues could be relevant for term-weighting, suggesting that structural features in particular would be worth pursuing. The intuition here is that keywords will tend to occur in specific places in the meeting, perhaps correlating to

topics or to speaker turns, whereas less informative words should occur more evenly throughout the meeting discussion. We structure the problem in the same manner that we did with the speaker probabilities; we segment the meeting into speaker turns, and for a given term t calculate the probability of each speaker turn T given t. We again take the maximum probability as our score Sc_2.

$$Sc_2(t) = \max_T p(T|t)$$

Finally, we hypothesize that term co-occurrence statistics are relevant to term-weighting for this data. Certain highly informative words should tend to occur together, whereas less informative words will have less discernible co-occurrence patterns. For example, "remote" and "control" may often appear together, as might "LCD" and "interface." For each term t, we identify its co-occurring terms based simply on the other word types that occur in the same dialogue acts as term t, after removal of stopwords. For each co-occurring term t_2, we calculate the ratio of the times t and t_2 co-occur to the total number of times that t_2 occurs in the meeting. So if the term t is "remote" and t_2 is "control" and out of the 20 times that "control" appears it co-occurs with "remote" 15 times, the score for t_2 is 0.75. We then take the maximum of all the co-occurring words' ratios as our score Sc_3.

$$Sc_3(t) = \max_t p(t|t_2)$$

We then combine the three scores Sc_1, Sc_2 and Sc_3 by calculating the harmonic mean of the scores. We hypothesize that exploiting such patterns in the meeting speech will be sufficient to carry out term-weighting for summarization, with no recourse to either collection frequency or term frequency. The only external source of information is the short stopwords list. We also stipulate that a term occurring 3 or fewer times in a meeting receives a score of 0. The reason is that a term that occurs only once or twice will have a very high score according to the first two sub-scores, since the probability of the speaker and the turn will approach 1.

For ease of reference, we subsequently refer to this second speech-specific metric as *twssd*, for "term-weighting for spontaneous spoken dialogues."

4.1.3 Experimental Setup

This section briefly overviews all of the term weighting approaches implemented, the corpora used, and the protocol for summarization evaluation. In addition to *tf.idf*, *su.idf*

and *twssd*, we also implement *idf*, *ridf* and *Gain* for comparison, with *idf* serving as a baseline metric. A hybrid approach combining the rankings of *tf.idf* and *su.idf* was implement in the hope that the two methods would be complementary, perhaps locating different types of informative terms. For all collection frequency measures, we use a collection of documents from the AMI, ICSI, Broadcast News and MICASE[1] corpora. This consists of 200 documents from the domains of broadcast news, scenario meetings, non-scenario meetings and lectures, which provide a balanced sampling of diverse speech genres. Each document represents a single lecture, meeting, or broadcast. Each term-weighting method is run on both manual and ASR transcripts. All documents in the collection are stemmed using the Porter stemmer (Porter, 1997).

For both the AMI and ICSI test set meetings, we also calculate human summarization performance according to the following method. For each annotator, we create a summary of 700 words length based on ranking their annotated dialogue acts from most-linked to least-linked and extracting until the length limit is reached. These summaries are then evaluated with weighted precision against the annotations of the remaining human judges. This is done for each annotator, and the scores are subsequently averaged to give a single human performance rating. For our evaluation, each term-weighting approach is used to create a brief summary of each test set meeting, and the resulting summaries are then evaluated. In each case we sum term-scores over dialogue acts to create scores for the dialogue acts, which are the summary extraction unit. Dialogue acts are ranked from most informative to least informative, and are extracted until a length of 700 words is reached. These summaries are then evaluated using the *weighted precision* metric originally introduced by Murray et al. (2006).

4.1.4 AMI Results

This section presents the weighted precision results for the AMI corpus test set. On manual transcripts, the best approach overall is *su.idf*, with an average weighted precision score of 0.66, followed by *Gain* with an average score of 0.64. The worst approaches are *tf.idf* and *idf*, with both being significantly worse than *su.idf* according to paired t-test ($p<0.05$). The *twssd* approach averages 0.62, outperforming *tf.idf* and *idf* despite using no term-frequency or collection frequency information.

On ASR transcripts, *su.idf* and *ridf* are the top term-weighting methods, each with an average of 0.67. The worst approaches are again *tf.idf* and *idf*, each with an average

[1] http://quod.lib.umich.edu/m/micase/

Meet	**idf**	**sidf**	**tw**	**tfidf**	**com**	**ridf**	**gain**	**H**
ES2004a	0.42	0.46	0.55	0.47	0.46	0.59	0.60	**0.67**
ES2004b	0.58	0.61	0.65	0.55	0.61	0.60	0.61	**0.83**
ES2004c	0.68	0.59	0.70	0.70	0.68	0.69	0.73	**0.58**
ES2004d	0.80	0.80	0.93	0.89	0.96	0.77	0.80	**1.03**
ES2014a	0.54	0.67	0.59	0.68	0.67	0.77	0.60	**0.82**
ES2014b	0.58	0.86	0.89	0.80	0.78	0.77	0.83	**0.80**
ES2014c	0.77	0.79	0.77	0.71	0.92	0.80	0.75	**1.21**
ES2014d	0.41	0.53	0.38	0.44	0.52	0.46	0.36	**0.63**
IS1009a	0.66	0.81	0.69	0.64	0.74	0.73	0.62	**1.16**
IS1009b	0.72	0.68	0.54	0.67	0.60	0.58	0.69	**1.15**
IS1009c	0.37	0.55	0.33	0.36	0.45	0.43	0.32	**0.72**
IS1009d	0.46	0.79	0.61	0.61	0.79	0.72	0.66	**1.10**
TS3003a	0.50	0.53	0.55	0.49	0.57	0.63	0.62	**0.68**
TS3003b	0.64	0.75	0.60	0.54	0.54	0.74	0.74	**0.98**
TS3003c	0.66	0.88	0.76	0.79	0.88	0.81	0.84	**0.94**
TS3003d	0.48	0.45	0.48	0.46	0.46	0.59	0.57	**0.70**
TS3007a	0.26	0.39	0.50	0.31	0.35	0.49	0.48	**0.86**
TS3007b	0.63	0.60	0.57	0.57	0.60	0.61	0.61	**0.65**
TS3007c	0.75	0.60	0.62	0.57	0.65	0.50	0.65	**0.92**
TS3007d	0.52	0.76	0.73	0.60	0.73	0.62	0.64	**0.86**
AVE:	**0.57**	**0.66**	**0.62**	**0.59**	**0.65**	**0.64**	**0.64**	**0.87**
Meet	**iasr**	**sasr**	**twasr**	**tfasr**	**casr**	**rasr**	**gasr**	**H**
ES2004a	0.47	0.60	0.55	0.61	0.63	0.66	0.61	-
ES2004b	0.68	0.54	0.63	0.59	0.56	0.65	0.60	-
ES2004c	0.67	0.69	0.87	0.71	0.67	0.74	0.68	-
ES2004d	0.66	0.75	0.82	0.81	0.83	0.74	0.73	-
ES2014a	0.53	0.73	0.71	0.69	0.74	0.84	0.72	-
ES2014b	0.78	0.87	0.80	0.74	0.80	0.74	0.70	-
ES2014c	0.77	0.78	0.82	0.64	0.88	0.76	0.70	-
ES2014d	0.38	0.45	0.48	0.45	0.46	0.40	0.46	-
IS1009a	0.77	0.94	0.72	0.73	0.78	0.83	0.69	-
IS1009b	0.70	0.67	0.55	0.57	0.65	0.72	0.59	-
IS1009c	0.32	0.53	0.33	0.40	0.44	0.49	0.47	-
IS1009d	0.57	0.70	0.64	0.61	0.67	0.64	0.64	-
TS3003a	0.49	0.59	0.56	0.49	0.60	0.58	0.54	-
TS3003b	0.74	0.74	0.59	0.59	0.68	0.71	0.76	-
TS3003c	0.70	0.89	0.66	0.68	0.82	0.86	0.89	-
TS3003d	0.55	0.54	0.48	0.51	0.52	0.60	0.52	-
TS3007a	0.40	0.51	0.55	0.45	0.51	0.50	0.51	-
TS3007b	0.60	0.54	0.60	0.51	0.52	0.70	0.63	-
TS3007c	0.62	0.61	0.57	0.57	0.64	0.58	0.57	-
TS3007d	0.66	0.71	0.75	0.63	0.64	0.63	0.57	-
AVE:	**0.60**	**0.67**	**0.63**	**0.60**	**0.65**	**0.67**	**0.63**	-

Table 4.2: Weighted Precision Results for AMI Test Set Meetings, Manual and ASR Transcripts

idf=*idf* on manual, **iasr**=*idf* on ASR, **sidf**=*su.idf* on manual, **sasr**= *su.idf* on ASR, **tw**=*twssd* on manual, **twasr**=*twssd* on ASR, **tfidf**=*tf.idf* on manual, **tfasr**=*tf.idf* on ASR, **com**=combined *su.idf* and *tf.idf* on manual, **casr**=combined *su.idf* and *tf.idf* on ASR, **ridf**=*ridf* on manual, **rasr**=*ridf* on ASR, **gain**=*Gain* on manual, **gasr**=*Gain* on ASR, **H**=human performance

Meet	Summ-WER	NonSumm-WER
ES2004a	34.7	49.9
ES2004b	30.9	38.6
ES2004c	27.5	39.0
ES2004d	34.0	46.6
ES2014a	40.5	50.0
ES2014b	32.1	47.1
ES2014c	33.1	47.4
ES2014d	32.6	46.8
IS1009a	32.0	41.0
IS1009b	30.2	36.8
IS1009c	38.8	39.6
IS1009d	29.6	38.0
TS3003a	25.2	44.4
TS3003b	22.6	28.5
TS3003c	21.6	33.0
TS3003d	24.2	35.2
TS3007a	30.6	41.5
TS3007b	22.8	35.3
TS3007c	30.7	40.2
TS3007d	32.5	41.9
AVERAGE	**30.31**	**41.04**

Table 4.3: Word Error Rates for Extracted (**Summ-WER**) and Non-Extracted Portions (**NonSumm-WER**) of Meetings, using *su.idf*

of 0.61. Each of *tf.idf* and *idf* are significantly worse than each of *su.idf*, *ridf* and the combined approach (all $p<0.05$).

With the exception of *Gain*, every term-weighting method improves on ASR compared with manual transcripts. Table 4.2 gives results on both manual and ASR.

It is particularly surprising that nearly all of the term-weighting approaches perform better on ASR than on manual transcripts. Previous research (Valenza et al., 1999; Murray et al., 2005a) has shown that informative portions of speech data tend to have lower word-error rates, but it is nonetheless unexpected that weighted precision would actually *improve* on errorful ASR transcripts. The *ridf* and *su.idf* metrics are particularly resilient to the errorful transcripts on this test set. Table 4.3 shows the word-error rates for the extracted and non-extracted portions of meetings using the *su.idf* summarizer. The WER for the extracted portions is more than 10 points lower than for the non-extracted portions of meetings, at 30.31% versus 41.04%. The WER for the corpus as a whole is around 38.9% for this particular version of the ASR. It should be noted that ASR word errors within a given summary do not contribute to a lower evaluation score, since our evaluation works at the dialogue act level; the im-

pact of the ASR errors is in determining whether or not a dialogue act is chosen for extraction in the first place.

It is also very encouraging that for several of the AMI test set meetings, the best automatic summarizers perform equal to or above human performance for weighted precision for this length summary. The average human performance across the test set, however, is still considerably higher than the performance of the automatic methods.

4.1.5 ICSI Results

This section presents the weighted precision results for the ICSI corpus test set. On both manual and ASR transcripts there were fewer differences between term-weighting approaches than we find on the AMI test set. On manual transcripts, the highest-scoring term-weighting schemes on average according to weighted precision are *Gain* and *twssd* with 0.37 each. The worst scoring method overall is *idf*, which averages 0.30. The only significant differences between all approaches are *idf* being significantly worse than *Gain* ($p<0.05$) and *twssd* ($p<0.1$).

On ASR, there are again few differences between all of the approaches, but within the individual weighting schemes there are some interesting differences between using manual and ASR transcripts. The highest scoring method on average is *ridf* with a score of 0.42, followed by *tf.idf* and *Gain*. The worst overall is the combined method of *tf.idf* and *su.idf*, with the only significant result being that this combined method is significantly worse than *ridf*.

Interestingly, *idf* performs much better on ASR than on manual transcripts for the ICSI corpus. With manual transcripts it is significantly worse than the top two term-weighting schemes, but increases seven points when applied to ASR. We also find that no single weighting scheme performs worse on ASR compared with manual; the precision results either remain the same or improve.

As can be seen in Table 4.4, the weighted precision scores in general are much lower than on the AMI meetings. However, the human performance is also much lower, illustrating low inter-annotator agreement for the ICSI corpus. In fact, the automatic summarization methods presented here perform *better* on the ICSI corpus than on the AMI corpus, by comparison with human gold-standard summarization. The best automatic methods are near or at the level of human summarization for this test set.

Meet	idf	sidf	tw	tfidf	com	ridf	gain	H
Bed004	0.21	0.19	0.27	0.31	0.19	0.26	0.32	**0.41**
Bed009	0.37	0.43	0.35	0.44	0.42	0.43	0.42	**0.39**
Bed016	0.36	0.33	0.43	0.39	0.48	0.48	0.41	**0.42**
Bmr005	0.34	0.41	0.39	0.27	0.36	0.46	0.41	**0.52**
Bmr019	0.17	0.37	0.37	0.29	0.33	0.20	0.33	**0.40**
Bro018	0.36	0.33	0.42	0.32	0.33	0.35	0.35	**0.34**
AVE:	**0.30**	**0.34**	**0.37**	**0.34**	**0.35**	**0.36**	**0.37**	**0.41**
Meet	**iasr**	**sasr**	**twasr**	**tfasr**	**casr**	**rasr**	**gasr**	**H**
Bed004	0.23	0.27	0.27	0.32	0.28	0.26	0.35	-
Bed009	0.45	0.37	0.37	0.44	0.35	0.48	0.39	-
Bed016	0.51	0.44	0.42	0.50	0.44	0.59	0.45	-
Bmr005	0.30	0.44	0.40	0.41	0.34	0.46	0.41	-
Bmr019	0.28	0.33	0.38	0.37	0.35	0.32	0.39	-
Bro018	0.41	0.33	0.36	0.38	0.36	0.42	0.35	-
AVE:	**0.36**	**0.36**	**0.37**	**0.40**	**0.35**	**0.42**	**0.39**	-

Table 4.4: Weighted Precision Results for ICSI Test Set Meetings, Manual and ASR Transcripts

idf=*idf* on manual, **iasr**=*idf* on ASR, **sidf**=*su.idf* on manual, **sasr**= *su.idf* on ASR, **tw**=*twssd* on manual, **twasr**=*twssd* on ASR, **tfidf**=*tf.idf* on manual, **tfasr**=*tf.idf* on ASR, **com**=combined *su.idf* and *tf.idf* on manual, **casr**=combined *su.idf* and *tf.idf* on ASR, **ridf**=*ridf* on manual, **rasr**=*ridf* on ASR, **gain**=*Gain* on manual, **gasr**=*Gain* on ASR, **H**=human performance

4.1.6 Weighted Recall and F-Score

The results in the sections above are solely weighted precision results, without recall or f-score information. The reasons for that are two-fold. The first is a somewhat historical reason, as the original formulation of the weighted evaluation was simply weighted *precision* and the initial term-weighting work of Murray and Renals (2007) used only that metric. The second and main reason is that the summaries are quite brief and are of equal lengths, and so recall scores across the board are very low. Since the intention is to create very concise and informative summaries and not to extract every relevant dialogue act, weighted precision is of much higher interest. For completeness, we present the weighted recall and f-score averages here, with Table 4.5 showing the AMI results and Table 4.6 showing the ICSI results. In later chapters we derive weighted f-scores for completeness as well.

Further analysis is provided in Appendix F (page 191), where we present precision and recall scores for the intersection and union of human-selected dialogue acts.

	idf	iasr	sidf	sasr	tw	twasr	tfidf	tfasr	com	casr	ridf	rasr	gain	gasr	H
R	0.16	0.16	0.18	0.18	0.18	0.19	0.18	0.18	0.18	0.18	0.18	0.18	0.16	0.16	**0.35**
F	0.23	0.24	0.27	0.27	0.26	0.27	0.26	0.27	0.27	0.27	0.26	0.26	0.24	0.24	**0.47**

Table 4.5: Weighted Recall and F-Score Averages for AMI Test Set

idf=*idf* on manual, **iasr**=*idf* on ASR, **sidf**=*su.idf* on manual, **sasr**= *su.idf* on ASR, **tw**=*twssd* on manual, **twasr**=*twssd* on ASR, **tfidf**=*tf.idf* on manual, **tfasr**=*tf.idf* on ASR, **com**=combined *su.idf* and *tf.idf* on manual, **casr**=combined *su.idf* and *tf.idf* on ASR, **ridf**=*ridf* on manual, **rasr**=*ridf* on ASR, **gain**=*Gain* on manual, **gasr**=*Gain* on ASR, **H**=human performance

	idf	iasr	sidf	sasr	tw	twasr	tfidf	tfasr	com	casr	ridf	rasr	gain	gasr	H
R	0.08	0.10	0.08	0.09	0.09	0.10	0.10	0.12	0.09	0.09	0.09	0.11	0.08	0.09	**0.14**
F	0.12	0.15	0.13	0.15	0.14	0.15	0.15	0.19	0.14	0.15	0.14	0.17	0.13	0.15	**0.20**

Table 4.6: Weighted Recall and F-Score Averages for ICSI Test Set

idf=*idf* on manual, **iasr**=*idf* on ASR, **sidf**=*su.idf* on manual, **sasr**= *su.idf* on ASR, **tw**=*twssd* on manual, **twasr**=*twssd* on ASR, **tfidf**=*tf.idf* on manual, **tfasr**=*tf.idf* on ASR, **com**=combined *su.idf* and *tf.idf* on manual, **casr**=combined *su.idf* and *tf.idf* on ASR, **ridf**=*ridf* on manual, **rasr**=*ridf* on ASR, **gain**=*Gain* on manual, **gasr**=*Gain* on ASR, **H**=human performance

4.1.7 Discussion

There are several interesting and surprising results from the experiments above. Perhaps the most surprising is that some of the metrics, especially *su.idf* and *ridf*, are particularly resilient to ASR errors, and we find a general trend that weighted precision actually increases on ASR. On the ICSI corpus, all of the term-weighting approaches stay the same or do slightly better, while on the AMI corpus all metrics but *Gain* stay the same or improve. It may be that informative words also tend to be less confusible words.

We also found that most of our metrics easily outperform these implementations of the classic *idf* and *tf.idf* term-weighting schemes, with *su.idf*, *twssd* and *ridf* consistently performing the best. We found that while *su.idf* performs better on the AMI corpus than the ICSI corpus, the reverse is true for *twssd*. This may be due to the fact that *su.idf* relies mostly on differing word usage between speakers, while *twssd* incorporates other pieces of information such as structural cues and co-occurrence information. As described above, the AMI meetings are scenario meetings with well-defined roles such as *project manager* and *marketing expert*, whilst roles in the ICSI corpus are much less clearly defined. Because roles are associated with certain vocabularies (e.g. the marketing expert being more likely to say “trend” or “survey” than the others), perhaps it would be expected that *su.idf* would perform better on those meetings

than on meetings where roles are more opaque and the dynamics between participants are more informal.

A possible explanation for why there are fewer differences overall between term-weighting approaches on the ICSI corpus versus the AMI corpus is that the keywords in the ICSI meetings are simply more easily discerned due to their technical nature. The ICSI meeting discussions concern speech and language processing issues in addition to other computational topics, and words like 'markov' and 'construal,' which are frequent in the meetings and very rare in the document collection, will be correctly assigned high scores even by *idf* and *tf.idf*. In contrast, AMI meeting keywords tend to be words like 'remote,' 'survey' and 'price,' which are not all that rare in the document collection and may incorrectly receive low scores from those metrics as a consequence.

One general result is that *tf.idf* is not as sensitive to term importance as the other metrics, the only exception being the ICSI ASR scores. It seems telling then that it is also the only metric that weights a term highly for occurring frequently within the given document. It is perhaps too blunt, favoring a few terms by scoring them highly and scoring the others dramatically lower, leading to a severely limited view of importance within the meeting. A strength of *su.idf* is that a term need not be very frequent within a document nor very rare across documents in order to receive a high score.

In an attempt to understand why *su.idf* outperforms *tf.idf* on the AMI meetings and why it performs better on ASR than on manual transcripts, the correlation between term rank and term score is examined for both term-weighting schemes. Figure 4.1 shows the normalized term-weighting scores for an example meeting TS3003c, with the x-axis representing term rank in descending order. It can be seen that *tf.idf* scores a handful of words very highly and the rest of the words in the meeting have sharply lower scores. In contrast, the *su.idf* scores arranged by rank exhibit a more gradual decline. This phenomenon holds across the entire test set: for the 20 AMI test set meetings, the top keyword for each meeting according to *tf.idf* is on average scored 3.88 times higher than the keyword ranked 20th for the same meeting, whereas the top *su.idf* keyword is only 2.23 times higher than the 20th on average.

On ASR, this phenomenon is even more pronounced. As can be seen in Figure 4.2, *su.idf* scores an even greater number of words highly, while there are only a small handful of high-scoring *tf.idf* keywords. It seems that *tf.idf* causes a few terms to dominate the others, while *su.idf* performs smoother scoring. A possible explanation for the steeper *tf.idf* score/rank curves exemplified in Figures 4.1 and 4.2 is that word

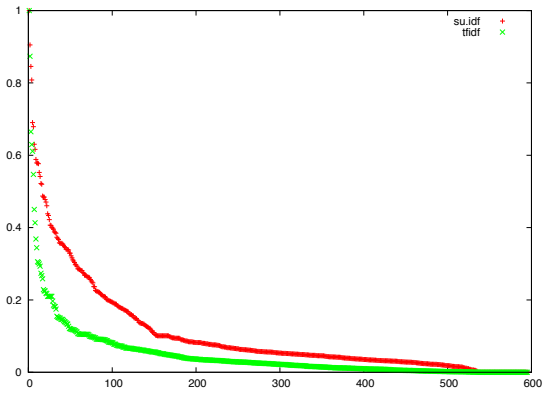


Figure 4.1: *Term Rank Plotted Against Term Score, Manual Transcripts*

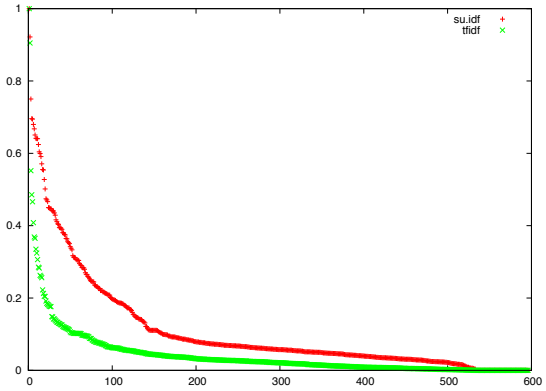


Figure 4.2: *Term Rank Plotted Against Term Score, ASR Transcripts*

frequency generally behaves according to the *zipf* distribution (Zipf, 1935), so that the n-most frequent word occurs approximately $\frac{1}{n}$ times as often as the most frequent word. Because term-frequency decreases rapidly as rank decreases, and *tf.idf* has a term-frequency component, perhaps it's not surprising that a few terms that occur most often score much higher than the rest of the terms. On this particular meeting, *su.idf* far outscores *tf.idf* according to weighted precision, on both ASR and manual transcripts.

Taking this analysis into account, the reason why *tf.idf* scores lower as far as weighted precision may be that because the weighting scheme favors words that occur many times in a document, there is less certainty about *which* dialogue acts to extract. If the top-scoring term occurs 30 times in a meeting, it's not clear which dialogue acts featuring that term should be extracted. As noted in section 4.1.1, metrics that favor mid-frequency terms have a history of performing well, and *su.idf* favors mid-frequency terms generally. Additionally, there are simply *more* high-scoring words, so that a meeting is not dominated by only 2 or 3 high-scoring terms. Using a weighting

scheme such as this implementation of *tf.idf* may increase summary redundancy by extracting many dialogue acts containing the same small set of high-scoring words.

To more closely inspect the differences in how each term-weighting scheme ranks words, a second evaluation for the same sample meeting is devised. Terms are ranked from highest to lowest; beginning with the highest-scoring term and proceeding until the hundredth highest-scoring term, all dialogue acts including that term are extracted and precision/recall/f-score are calculated. Then each dialogue act containing the first or second terms are extracted, and precision/recall/f-score are again calculated. This continues down the top 100 list of terms, with precision/recall/f-score calculated at each step.

Figure 4.3 shows precision, recall and f-score for ranked *tf.idf* scores on meeting TS3003c, while Figure 4.4 shows the same measures for ranked *su.idf* scores. Unsurprisingly, the *tf.idf* recall scores initially rise more sharply - because the metric favors terms that occur often in a meeting, more dialogue acts are extracted at first. However, *tf.idf* suffers in terms of precision compared to *su.idf*.

To give an example of the effect of *tf.idf* on actual summarization output, we again consider meeting TS3003c. In this meeting, the top terms according to *tf.idf* are "remote," "button," "docking," "subtitle," "and trend-watching." These five words score so highly compared with all other terms in the meeting that every single dialogue act selected contains one of the five words. These results in a high level of redundancy, as evidenced by this excerpt:

> **Speaker D:** Uh the remote control and the docking station should uh blend in in the in the room.
>
> **Speaker D:** Um well the trend-watchers I consulted advised that it b should be, the remote control and the docking station should be telephone-shaped.
>
> **Speaker D:** So you could imagine that uh the remote control will be standing up straight in the docking station.
>
> **Speaker D:** So they would prefer uh a design where the remote control just lies flat in the docking station.

Taking all of these findings together, it seems that *su.idf* succeeds by not necessarily favoring frequent words in a meeting, by not scoring a word with a low weight simply because it is fairly common in other documents, and by having a smooth scoring curve rather than the steep drop-off in scores found with *tf.idf*. The metric *tf.idf* tends to have a very small handful of words dominate, thereby skewing the extraction process to favor a handful of similar dialogue acts. The metric *idf* alone performed

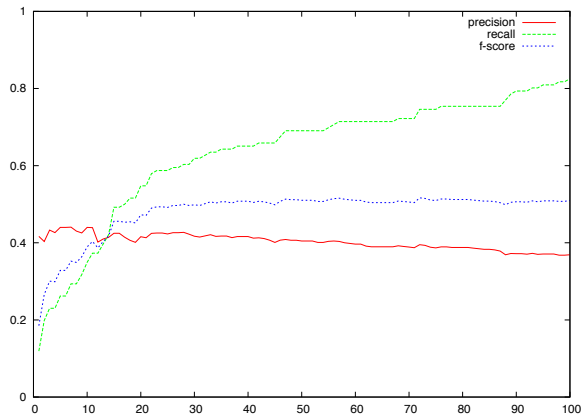


Figure 4.3: *Precision/Recall/F-Score, TFIDF, top 100 keywords*

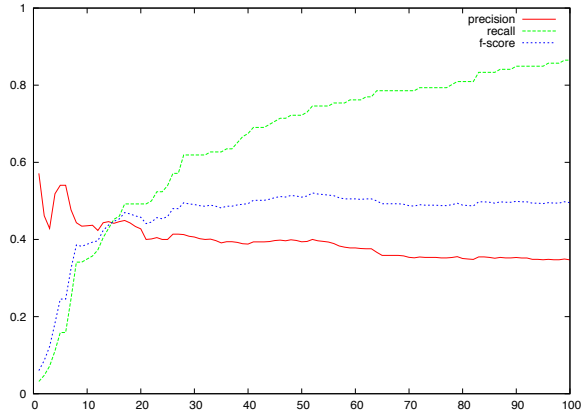


Figure 4.4: *Precision/Recall/F-Score, SUIDF, top 100 keywords*

worst of all, illustrating that a term being rare across documents is by itself not necessarily indicative of its informativeness for summarization purposes.

4.1.8 Term-Weighting Conclusion

We have presented an evaluation of term-weighting metrics for spontaneous, multi-party spoken dialogues. Four of the metrics, *idf*, *tf.idf*, *ridf* and *Gain*, were imported from text IR to test for suitability with our data. Two novel approaches called *su.idf* and *twssd* were implemented, the former relying on the differing patterns of word usage among meeting participants, with the latter also including structural and co-occurrence information. Both were found to perform very competitively, with *su.idf* scoring very highly on the AMI data and *twssd* scoring highly on the ICSI data. The other major findings are that all the term-weighting techniques investigated are fairly robust to ASR errors, and that it is easy to outperform the standard implementations of

idf and *tf.idf* baselines on this type of data.

4.2 Cue Words for Summarization

This section examines how cuewords can be used for summarization purposes, and how summarization using only cueword information compares to the summarization results described in the previous section.

Cuewords are words that signal informativeness or areas of interest but are not specific to the topic being discussed, unlike keywords. For example, words such as "important" or "decide." Since these words are relatively common across documents, they will normally not be rated highly by a term-weighting scheme such as those described and implemented in section 4.1. The hypothesis is that they are a valuable source of information for summarization and that using only keyword detection and not cueword detection causes us to miss important dialogue acts.

As mentioned in the introduction, one of the interesting findings in the seminal work of Edmundson (1969) is that cuewords are often as good or better than keywords for the purpose of automatic summarization. This section examines what level of summarization performance we can attain if we use *only* cuewords and none of the term-weighting schemes described in section 4.1. If we can maintain comparatively high precision results without using term-weighting metrics and instead focusing on the presence of certain trigger words, this will be very useful when conducting online or real-time summarization; in such scenarios, full speech recognition output may be either unavailable or very degraded, in which case we can utilize limited cueword spotting in place of full recognition. Such online meeting analysis is the central focus of the AMIDA project[2] and online keyword and cueword spotting are essential to further analyses.

4.2.1 Determining Cuewords

We are interested in which terms are likely to signal informative dialogue acts, and so we conduct a corpus analysis of term-frequencies between those dialogue acts labelled as "extractive" and those labelled as "non-extractive" in the training data. To begin, a list of around 200 potential cuewords is constructed, comprised of terms that would be expected to signal informative dialogue acts in a meeting environment, based on

[2] http://www.amiproject.org

introspection. These are terms that are not specific to the AMI scenario, so words such as "remote" are excluded. Examples of words on the original list are "important," "decide", "discuss" and "group" - words that merit inclusion based on general intuitions about meeting dialogue and which are not specific to individual meetings. For each word on this initial list, we then compare its normalized frequency in extract portions of meetings to its normalized frequency in non-extracted portions of meetings in the training data, and score each term thusly,

$$TF(t,j)/TF(t,k)$$

where $TF(t,j)$ is the number of times that term t occurs in the extracts normalized by the total number of tokens in the extracts, and $T(t,k)$ is the number of times that term t occurs in the non-extracts normalized by the total number of tokens in the non-extracts. We discard any terms on the original list that do not occur at least 50 times in the training data, so as not to be thrown off by small sample sizes.

Ranking the words in the list according to this ratio, we then keep the top 70 words and discard the remainder. This was done for both the AMI and ICSI data, on both manual and ASR transcripts, for a total of four unique cueword lists. The cutoff of the top 70 words was chosen because the trend seemed to be that the ratio reached 1 between the 80th and 90th positions of the list, where there was no longer any differences between extract and non-extract frequencies of occurrence for the remainder of the items on the list.

Appendix B (page 179) lists the four cuewords lists in their entirety. We include here the top 10 cuewords for each of the four lists in Table 4.7. Note that these words represent stems, and so, for example, "expect" will match "expect", "expectation", "expected", etc.

We restrict ourselves to unigram cues here, but one could compare n-gram occurrences between extracts and non-extracts at higher values of n (Galley, 2006).

4.2.2 Cueword-Based Summarization

Summarization proceeds as before, now with each sentence scored by summing over its constituent cueword scores. If a term does not exist on the relevant cuewords list, its score is simply zero. We then rank the sentences and extract the best sentence in turn until we reach our desired length - in this case, 700 words.

Rank	AMI-MAN	AMI-ASR	ICSI-MAN	ICSI-ASR
1	expect	expect	focus	focus
2	found	component	fairly	soon
3	component	found	area	fairly
4	project	fairly	group	apparent
5	focus	agenda	project	study
6	group	focus	report	report
7	research	project	soon	group
8	meet	group	decision	project
9	final	research	topic	finish
10	agenda	team	summarize	response

Table 4.7: Top 10 Cuewords, AMI and ICSI, Manual and ASR

4.2.3 Results

Table 4.8 gives the weighted precision scores for the AMI test set cueword-based summaries. The average weighted precision score for this method applied to manual transcripts is 0.55 while the average weighted precision score for ASR transcripts is 0.53. In contrast, the weighted precision results for *tf.idf* reported above were 0.60 for both transcript types. So we find that while the cueword approach does not perform as well as the keyword-based approaches of section 4.1, its scores are not dramatically lower despite using a very small vocabulary of cuewords and no keyword information for the documents to be summarized. While it is clearly advantageous to use keyword information, it is not strictly necessary for generating good-quality summaries, and in situations where full speech recognition is not available, limited cueword spotting using these lists of terms would suffice for indicating areas of high informativeness in the meeting.

It also seems to be the case that keyword summaries and cueword summaries are complementary in some respects. There are several meetings on which the keyword approaches do not score highly and the cueword approach is considerably better, e.g. IS1009b and IS1009c. On those two meetings in particular, the cuewords comprise a higher-than-average percentage of the total word tokens in the meeting, at about 10% in each meeting. This may explain the high quality of the cueword summaries on those meetings. In contrast, the cueword summary for TS3007c is substantially lower than the keyword summaries for the same meeting, and the cueword form only around 7% of tokens in that meeting. These findings suggest that a larger cuewords list could substantially increase summarization performance, as we are liable to find, when using a vocabulary of only 70 items, that some meetings will simply not contain

Meet	CUE-Man	CUE-ASR
ES2004a	0.44	0.46
ES2004b	0.49	0.53
ES2004c	0.62	0.57
ES2004d	0.7	0.65
ES2014a	0.52	0.43
ES2014b	0.39	0.39
ES2014c	0.6	0.56
ES2014d	0.3	0.34
IS1009a	0.65	0.66
IS1009b	1.01	0.91
IS1009c	0.4	0.37
IS1009d	0.5	0.58
TS3003a	0.42	0.32
TS3003b	0.78	0.68
TS3003c	0.67	0.58
TS3003d	0.37	0.31
TS3007a	0.47	0.49
TS3007b	0.5	0.56
TS3007c	0.47	0.47
TS3007d	0.73	0.65
AVERAGE	**0.55**	**0.53**

Table 4.8: AMI Corpus, Weighted Precision Scores on Manual and ASR Transcripts

many of those cues. For meetings that do contain a fair number of these cuewords, summarization performance is credible.

It is also encouraging that this approach, like the keyword approaches, is resilient to the ASR errors in the transcript. Its average on ASR is only slightly lower, and in many cases a given meeting summary is higher on ASR than on manual transcripts. This is somewhat surprising – while it is well-known that summarizers are often resilient to ASR errors (Valenza et al., 1999; Murray et al., 2005a), the lack of degradation is often attributed to keywords being less confusible words. Keywords tended to be longer and more technical, which is not the case with many of the cuewords used here. Nonetheless performance on ASR is robust. The WER for the AMI cuewords summaries is 29.8%, which is even slightly lower than for the AMI *su.idf* summaries.

Table 4.9 gives the weighted precision results on the ICSI corpus. The averages on manual and ASR transcripts are both 0.30, compared with *tf.idf* scores of 0.39 and 0.40 respectively from section 4.1. The gap between keyword and cueword scores is larger than with the AMI corpus scores, but the results nonetheless contain encouraging findings: that using only cuewords we can generate summaries of acceptable quality, and that this cuewords technique is resilient to ASR errors. A likely reason for

Meet	CUE-Man	CUE-ASR
Bed004	0.22	0.23
Bed009	0.38	0.34
Bed016	0.40	0.41
Bmr005	0.27	0.28
Bmr019	0.23	0.23
Bro018	0.32	0.27
AVERAGE	**0.30**	**0.30**

Table 4.9: ICSI Corpus, Weighted Precision Scores on Manual and ASR Transcripts

the cuewords approach performing better on the AMI data than the ICSI data is that the cueword density varies substantially between the two corpora. For the AMI data, the cuewords represent about 9% of all tokens in the test set meetings, whereas in the ICSI data the cuewords represent only 5.5% of tokens in the test set meetings. Apparently because the ICSI meetings are generally less structured than the AMI meetings, they have fewer structural cues such as discourse markers or other cuewords. This difference between cuewords-based summarization on AMI versus ICSI data again illustrates that a larger cuewords vocabulary would likely increase performance in general.

4.2.3.1 Weighted Recall and F-Score

While the precision results for the cuewords summaries are below the precision scores for the keyword summaries described earlier, the weighted recall scores are comparable and thus the weighted f-scores are only slightly lower. For the AMI corpus test set, the weighted recall and f-score with manual transcripts are 0.17 and 0.25, respectively, and for ASR transcripts they are 0.16 and 0.23. For the ICSI corpus test set, the weighted recall and f-score with both manual and ASR transcripts are 0.08 and 0.13 .

4.2.4 Cuewords Conclusion

This section has presented the findings of an experiment indicating that summarization using cueword detection can approach levels of summarization using term-weighting keyword detection. This finding is particularly relevant for situations where full speech recognition is unavailable and one must rely instead on keyword spotting. Our lists of informative cuewords can be used in such scenarios.

Although the weighted precision results are lower than summarization using keyword weights, these cueword summaries are generated based on a very small vocabu-

lary of cuewords consisting of only 70 word stems. Expanding this list could presumably increase summarization precision.

Chapter 7 (page 126) further investigates the use of cuewords, focusing on a particular type of cueword and incorporating that cueword information into a machine-learning framework. That chapter also detects cuewords from the training data in a fully automatic fashion, rather than beginning with a manually-written list of hypothesized cuewords that we subsequently refine.

4.3 Conclusion

This chapter has detailed experiments on term-weighting and cueword detection. Section 4.1 surveyed several term-weighting approaches and evaluated their usefulness on spontaneous speech data such as the AMI and ICSI corpora. This section also introduced two novel term-weighting metrics for multi-party spontaneous speech called *su.idf* and *twssd*, and found them to be competitive with the state-of-the-art.

Section 4.2 looked at the usefulness of cuewords for summarization, either as a supplement to keyword information or a replacement when a full ASR transcript is not immediately available. The finding was that cuewords alone are sufficient for the creation of good-quality summaries and that cuewords methods are resilient to ASR errors. Summaries created based on cueword detection can be further revised when full transcripts are available.

In subsequent chapters, both keywords and cuewords are utilized as features in a machine-learning framework.

Chapter 5

Extractive Summarization

In this chapter we examine the issue of extractive summarization and specifically investigate the most useful features for automatic extraction. For each meeting in the AMI and ICSI test sets, we aim to detect the most informative set of dialogue acts to extract in order to create a compression of the meeting as a whole. This chapter builds on Chapter 4 (page 38) regarding term-weighting, as term-weights are a useful feature for automatic summarization. However, this research aims to determine which additional features, particularly speech-specific features, are valuable for the extraction task, and whether approaches that use a variety of multi-modal features can outperform solely text-based approaches.

5.1 Extractive Summarization Overview

As mentioned in the introduction, our summarization paradigm is that of *extractive* summarization. Given a source document consisting of an ASR transcript and features derived from the speech signal, we want to detect which dialogue acts in the meeting are the most informative. The hypothesis here is that the optimal results will be found by using a combination of lexical, prosodic, structural and speaker-based features, rather than treating the problem as merely text summarization on noisy input data.

In the first section, we look at unsupervised approaches that are either taken directly from research in text summarization or are inspired by previous text summarization research. These methods are applied to the meeting transcripts, first using *tf.idf* and subsequently using *su.idf* as a comparison. This aims to establish how well unsupervised approaches can perform compared with machine-learning approaches incorporated many additional features, as well as comparing *tf.idf* and *su.idf* in more advanced

summarization systems than the simple system described in Chapter 4.

In the second section we describe the machine-learning approach to the summarization task, describing the logistic regression classifier and the features used. We present results based on weighted precision, recall and f-score and compare these results to the unsupervised methods. We also present an in-depth analysis of the individual features used and the classification performance of various feature subsets.

The summaries generated are 700 words in length. Because summarization performance can be linked to summary length, we also evaluate our extractive classifiers according to the receiver operator characteristic, which measures the true-positive/false-positive ratio of the test data, generalizing away from given posterior probability thresholds and particular summary lengths. This latter evaluation is the most comprehensive, as we derive ROC curves for various feature subsets in order to determine the most useful characteristics of the data for summarization purposes.

5.2 Importing Text Summarization Approaches

This section describes several text summarization approaches that were implemented for these experiments. The approaches are either well-known or are based on well-known principles within text summarization.

5.2.1 Maximal Marginal Relevance

Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998) is based on the vector-space model of text retrieval, and is well-suited to query-based and multi-document summarization. In MMR, sentences are chosen according to a weighted combination of their relevance to a query (or for generic summaries, their general relevance) and their redundancy with the sentences that have already been extracted. Both relevance and redundancy are measured using cosine similarity. The usual MMR score $Sc_{MMR}(i)$for a given sentence S_i in the document is given by

$$Sc_{MMR}(i) = \lambda(cos(S_i, q)) - (1-\lambda) \max_{S_j \in summ} (cos(S_i, S_j)),$$

where q is the query vector, *summ* is the set of sentences already extracted, and λ trades off between relevance and redundancy. The term *cos* is the cosine similarity between two documents. For these experiments, we use the general informativeness of a sentence as determined by the sum of its term-scores, rather than the similarity

$cos(S_i, q)$ of the sentence to an average document vector. For redundancy, we take the maximum cosine of the candidate sentence and each already-extracted sentence.

In our previous implementation of MMR (Murray et al., 2005a), the weight λ was annealed, so that relevance was emphasized when the summary was still short, and as the summary grew longer the emphasis was increasingly put on minimizing redundancy. For the first third of the summary, $\lambda = 0.7$, for the second third $\lambda = 0.5$, and for the final third of the summary $\lambda = 0.3$. In this implementation, we simply set λ at 0.7, as further experimentation is needed to prove the usefulness of λ annealing.

5.2.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a vector space approach which involves projection of the term-document matrix to a reduced dimension representation. It was originally applied to text retrieval (Deerwester et al., 1990), and has since been applied to a variety of other areas, including text summarization (Gong & Liu, 2001; Steinberger & Ježek, 2004). LSA is based on the singular value decomposition (SVD) of an $m \times n$ term-document matrix A, whose elements A_{ij} represent the weighted term frequency of term i in document j, where the document is a sentence. In SVD, the term-document matrix is decomposed as follows:

$$A = USV^T$$

where U is an $m \times n$ matrix of left-singular vectors, S is an $n \times n$ diagonal matrix of singular values, and V is the $n \times n$ matrix of right-singular vectors. The rows of V^T can be interpreted as defining topics, with the columns representing sentences from the document. Following Gong and Liu (2001), summarization proceeds by choosing, for each row in V^T, the sentence with the highest value. This process continues until the desired summary length is reached.

Steinberger and Ježek (2004) have offered two strong criticisms of the Gong and Liu approach. Firstly, the method described above ties the dimensionality reduction to the desired summary length. Secondly, a sentence may score highly but never "win" in any dimension, and thus will not be extracted despite being a good candidate for extraction.

We addresses the same concerns as Steinberger and Ježek, while still following the Gong and Liu approach. Rather than extracting the best sentence for each topic, the n best sentences are extracted, with n determined by the corresponding singular values

from matrix S. Thus, dimensionality reduction is no longer tied to summary length and more than one sentence per topic can be chosen. For each topic, the number of words to extract from that topic is equal to the ratio of the associated squared singular value and the sum of all squared singular values. For example, if the desired summary length is 1000 words, the square of its associated singular value is 16 and the sum of all squared singular values is 32, then 500 words are drawn from the first topic. The motivation for this is that the more important a topic is, the more it should be represented in the ultimate summary. Furthermore, the number of dimensions is learned from the data rather than explicitly supplied as with the Steinberger and Ježek method.

For these experiments we implement both the Steinberger/Ježek method and the novel approach described above.

5.2.3 Centroid Approaches

The third unsupervised method is a textual approach incorporating LSA into a centroid-based system (Radev et al., 2000, 2001). The centroid is a pseudo-document representing the important aspects of the document as a whole; in the work of Radev et al. (2000), this pseudo-document consists of keywords and their modified *tf.idf* scores. In the present research, we take a different approach to constructing the centroid and to representing sentences in the document. First, *tf.idf* scores are calculated for all words in the meeting. Using these scores, we find the top twenty keywords and choose these as the basis for our centroid. We then perform LSA on a very large corpus comprised of a concatenation of multiple speech corpora: the ICSI, AMI, Broadcast News, and MICASE corpora, supplemented by the much larger Acquaint news-wire corpus. We perform LSA on the data using the Infomap tool[1] (Widdows et al., 2003). Infomap operates by performing latent semantic analysis on a large term co-occurrence matrix, allowing us to derive underlying term similarities. Infomap provides a query language with which we can retrieve word vectors for our twenty keywords, and the centroid is thus represented as the average of its constituent keyword vectors (Foltz et al., 1998) (Hachey et al., 2005).

Dialogue acts from the meetings are represented in much the same fashion. For each dialogue act, the vectors of its constituent words are retrieved, and the dialogue act as a whole is the average of its word vectors. In previous experiments (Murray et al., 2006) using this LSA centroid representation, extraction proceeded simply by

[1] http://infomap.stanford.edu

measuring the cosine between the dialogue act vectors and the query vector. However, the centroid approach did not fare as well as other unsupervised methods in that set of experiments, and the extraction process is supplemented in the current work. It was hypothesized that relying solely on the LSA sentence representations for gauging similarity to the centroid may have caused dialogue acts to be extracted that were only vaguely related to the actual content of the centroid. That is, terms in a candidate dialogue may have had an underlying similarity to terms in the query despite the candidate dialogue act not being particularly informative. In an attempt to increase precision, we have therefore included two informativeness measures: the centroid similarity as calculated before, and a general informativeness score based on the sum of dialogue act term-scores.

Extraction then proceeds along the same lines as MMR, described above, with the harmonic mean of the two informativeness metrics as a single informativeness score, penalized by a redundancy metric that is the maximum cosine of the candidate sentence and all of the extracted sentences, using the LSA sentence vectors.

5.3 Speech-Specific Summarization Approaches

In this section we present summarization systems that exploit a variety of speech-specific characteristics, in contrast to the systems described in the previous section, which are entirely text-based.

5.3.1 Augmenting Text Summarization Approaches with SU.IDF

For each of the summarization approaches described in Section 5.2, which are imported from the field of text summarization, we run the systems with both *su.idf* and *tf.idf* as a further comparison of the term-weighting approaches incorporated into commonly-used summarizers. The hypothesis is that the extraction techniques themselves are transferable between domains but can be improved by modifying the term-weight inputs to reflect the speech data. The summarization systems presented in this chapter are more advanced than the relatively simple summarization method described previously in Chapter 4.

5.3.2 Feature-Based Approaches

As described in detail in Chapter 3 Section 3.5.2 (page 31), the classifier used for the following experiments is the *liblinear* logistic regression classifier.[2].

This section introduces the features that are used for the machine learning experiments and motivates their inclusion in the feature database.

The first class of features is prosodic features, or features having to do with suprasegmental characteristics of the speech signal. We take a "direct modeling" approach to prosody (Shriberg & Stolcke, 2004), deriving prosodic features directly from the signal rather than utilizing intermediate prosodic annotation schemes such as ToBI (Silverman et al., 1992) or RAP (Dilley et al., 2006). There are two energy features, mean energy and maximum energy. Both are normalized by speaker and by meeting, so as to cancel out effects of speaker variety and microphone proximity. The mean energy is the average energy level for the dialogue act, while the maximum energy feature is the maximum energy level for that dialogue act. Both of these feature are motivated by the observation that speakers tend to raise their voices in terms of intensity when they are engaged in heated discussion or emphasizing a particular point they believe to be salient.

Three F0 features are included in the features database: mean F0, max F0 and F0 standard deviation. For all F0 features, we discard the lower and upper 5th percentiles of F0 values for each speaker in each meeting, so as to exclude cases of pitch doubling or segments where the pitch tracker drops out. The mean F0 is the average F0 value for the dialogue act, the maximum F0 is the highest F0 value in the dialogue act, and the F0 standard deviation is a measure of the spread of the dialogue act's F0 values from the dialogue act mean. The perceptual correlate of F0 is pitch, and increases in pitch can often correlate with stress. Furthermore, a meeting participant speaking with an expressive pitch contour might be signalling increased engagement compared with a flat, monotone pronunciation. Of course, there are many factors that affect a speaker's pitch, such as emotion and syntactic structure, and these factors may outweigh or confound any pitch correlates of informativeness.

Other prosodic features relate to duration and pauses. The duration of the dialogue act in seconds is included because a longer dialogue act very likely contains more information than a short dialogue act. The length in number of words is also included, due to the fact that a dialogue act might be very long in terms of time duration but with

[2]http://www.csie.ntu.edu.tw/ cjlin/liblinear/

few informative words. A final duration feature is that of uninterrupted length, meaning the length in seconds of the portion of each dialogue act not overlapped by another dialogue act. This feature is included because it can signal areas of high multi-speaker interaction on the one hand, wherein participants speak on top of one another and interrupt each other, and areas where the current speaker clearly holds the floor without interruption. The duration features here are very dependent on dialogue act segmentation, although it's of course possible to use extraction units of varying granularity. In Chapter 8 Section 8.3 (page 165) we explore the use of speaker spurts as opposed to dialogue acts.

Two pause features are included, precedent pause and subsequent pause. Both are based on the idea of ordering the meeting dialogue acts monotonically according to start time. A dialogue act's precedent pause is then the difference between its start time and the end time of the preceding dialogue act. This value can therefore be negative, signalling dialogue act overlap. Similarly, subsequent pause is the difference between the start time of the subsequent dialogue act and the end of the current dialogue act, a value that can also be negative. These features clearly help to indicate areas of high interaction by signalling overlap, as well as indicating long pauses that may be due to a speaker gathering their thoughts in order to say something.

The final prosodic feature is a rough estimate of rate-of-speech, which is simply the number of words in the dialogue act divided by the duration of the dialogue act in seconds. A meeting participant speaking very rapidly might indicate that the speaker is engaged and conveying a large amount of information, and that that particular section of the meeting might be a region of high information density. There are other possible ways to measure rate-of-speech, many of them based on syllables or on voiced versus unvoiced frames, but here we choose an intuitive method that is quick to derive.

The features database includes two structural features. The first is the dialogue act's position in the meeting, with 0 representing the beginning of the meeting and 1 representing the end of the meeting. This feature is somewhat domain-specific, as it would be expected that the beginnings and ends of meetings would have many information-rich utterances due to the meeting leader introducing the topics to be discussed at the beginning or summarizing what was discussed at the end.

The second structural feature is the dialogue act's position in the speaker's turn. For example, a speaker might speak four dialogue acts in a row, in which case the position of a given dialogue act among those four might be significant. Perhaps the speaker is building to a point; or conversely, expanding on the initial point. A further

motivation for including structural features in the database is that structural features on textual data are used with great success (Edmundson, 1969) – for example, the position of a sentence within an article and within a paragraph – and we similarly hope to exploit structural characteristics of a speech record. Of course, speech data inherently has less structure than text and so there are fewer structural characteristics to exploit in comparison. Further meeting structure could possibly be exploited, such as using automatic topic segmentation output (Hirschberg & Nakatani, 1998; Hsueh & Moore, 2006), but in this research we concentrate on more easily derived meeting structure.

The next class of features is related to speaker status, and these features aim to measure how dominant a given speaker is in a meeting. The first such feature is speaker dominance according to number of dialogue acts spoken; specifically, what percentage of total dialogue acts in the meeting does the speaker of the given dialogue act account for? The second feature is similar but measures the dominance in speaking time rather than number of dialogue acts: what percentage of total meeting speaking time does the speaker of the given dialogue act account for? These dominance features are somewhat domain-specific, and are included based on the intuition that a person who is more dominant in a meeting is a person of higher status in the meeting group, e.g. the project manager, and that such a person is more likely to utter high-level informative utterances relating to the topics and agenda of a meeting. Such a speaker is also more likely to summarize topics and the meeting as a whole. It should be noted that these are not features of the individual dialogue acts, but rather features of the speakers of the dialogue acts. The idea of dominance here is also fairly limited, as it relates purely to social dominance and floor-holding; these features do not incorporate ideas of specific participant roles, a speaker's influence on other speakers, or deference to a particular individual, to give a few examples. It is also possible that a given speaker may be dominant on a particular issue or topic but not in the meeting as a whole.

The final class of features in this database is lexical features of informativeness. The two features of this class are *tf.idf* and *su.idf*, described in detail in Chapter 4 (page 38). These features are included because it is presumed that informative dialogue acts will tend to contain words with high term-weighting scores, and that the two weighting methods may be complementary. Term-weight features in general are also motivated by having been used with success in previous summarization work on both text and speech. Each feature represents the sum of term scores across the given dialogue act.

To summarize the feature database overall, we have included features of prosody,

Feature ID	Description
ENMN	mean energy
F0MN	mean F0
ENMX	max energy
F0MX	max F0
F0SD	F0 stdev.
MPOS	meeting position
TPOS	turn position
DDUR	d. act duration
PPAU	precedent pause
SPAU	subsequent pause
UINT	uninterrupted length
WCNT	number of words
DOMD	dominance (d. acts)
DOMT	dominance (seconds)
ROS	rate of speech
SUI	su.idf sum
TFI	tf.idf sum

Table 5.1: Features Key

meeting structure, speaker status, and lexical informativeness. They have been motivated by success in previous research, linguistic insight and intuitions on meeting dynamics and structure. With the exception of *tf.idf* and word-count, every feature captures a characteristic of the data that is specific to speech data. Table 5.1 summarizes the features used.

5.4 Evaluation Protocols

The work in this section relies entirely on weighted precision/recall/f-score for evaluation. However, Chapter 6 (page 93) describes a large-scale extrinsic evaluation for a variety of summary types, and Chapters 7 (page 126) and 8 (page 148) incorporate ROUGE as an evaluation metric for specialized purposes.

5.5 Results - Imported Unsupervised Methods

Section 5.2 (page 64) described five unsupervised summarization methods that were applied to our speech data, using *su.idf* and *tf.idf* as term-vector weights, as well as running on both manual and ASR transcripts. The following sections report the results on both AMI and ICSI meeting data.

5.5.1 AMI Results

For all of the unsupervised methods applied to the AMI test data, the LSA centroid method is superior on both manual and automatic transcripts when using the *tf.idf* term-weighting scheme. The average f-score is 0.26 for manual transcripts and 0.27 for automatic transcripts (the f-scores seem somewhat low because recall is very low due to the short summary length). The average for MMR, the Steinberger/Ježek approach and the novel SVD approach are 0.20, 0.20 and 0.18 for manual transcripts respectively, and 0.19, 0.19 and 0.18 respectively when applied to ASR. The LSA centroid method is the only unsupervised method for *tf.idf* that shows slight improvement on ASR, though the other approaches do not show marked degradation. For both manual and ASR, the LSA centroid method is significantly better than each of the other methods according to paired t-test (all $p<0.05$). Table 5.2 shows the weighted f-scores for each meeting on both manual and ASR transcripts.

When using *su.idf* as the term-weighting scheme, the unsupervised approaches in general show significant improvement in terms of weighted f-scores compared with *tf.idf*. On manual transcripts, MMR improves from 0.20 to 0.25, the Steinberger/Ježek approach improves from 0.20 to 0.29 and the novel SVD approach improves from 0.18 to 0.27. The LSA Centroid method on manual transcripts is the same using both *su.idf* and *tf.idf*. The Steinberger/Ježek is significantly better than MMR and the LSA Centroid approaches according to paired t-test ($p<0.05$) and significantly better than the novel LSA approach ($p<0.10$). Table 5.3 shows the weighted f-score results for each meeting.

On ASR transcripts using *su.idf*, the LSA centroid method and the Steinberger/Jezek method are slightly superior. The LSA centroid method again improves slightly on ASR transcripts as compared with manual transcripts, with an average f-score of 0.28. It is significantly better than MMR according to paired t-test ($p<0.05$).

5.5.2 ICSI Results

On the ICSI data using *tf.idf* as the term-weighting scheme, the LSA centroid method is significantly better than the other approaches (all $p<0.05$), with an average weighted f-score of 0.13. All of the summarization methods show slight improvement when applied to ASR, with the LSA centroid method again performing best, with an average weighted f-score of 0.15. The centroid method is still significantly better than the Steinberger/Ježek method and the novel LSA approach (both $p<0.05$). Table 5.4 shows

Meet	MMR	CENTR	LSA-SJ	LSA-Murr
ES2004a	0.32	0.44	0.26	0.28
ES2004b	0.17	0.17	0.13	0.16
ES2004c	0.14	0.25	0.09	0.13
ES2004d	0.14	0.25	0.16	0.12
ES2014a	0.29	0.49	0.31	0.26
ES2014b	0.17	0.21	0.17	0.18
ES2014c	0.13	0.23	0.12	0.14
ES2014d	0.06	0.21	0.14	0.11
IS1009a	0.39	0.49	0.30	0.30
IS1009b	0.21	0.18	0.25	0.20
IS1009c	0.26	0.21	0.23	0.17
IS1009d	0.12	0.33	0.20	0.16
TS3003a	0.29	0.32	0.28	0.23
TS3003b	0.17	0.18	0.21	0.18
TS3003c	0.22	0.24	0.21	0.15
TS3003d	0.22	0.23	0.22	0.22
TS3007a	0.26	0.25	0.27	0.25
TS3007b	0.13	0.17	0.16	0.15
TS3007c	0.15	0.17	0.16	0.18
TS3007d	0.14	0.20	0.15	0.13
AVERAGE	**0.20**	**0.26**	**0.20**	**0.18**
ES2004a-ASR	0.27	0.41	0.24	0.23
ES2004b-ASR	0.17	0.22	0.12	0.15
ES2004c-ASR	0.18	0.21	0.15	0.14
ES2004d-ASR	0.17	0.27	0.21	0.22
ES2014a-ASR	0.32	0.53	0.35	0.33
ES2014b-ASR	0.13	0.23	0.13	0.13
ES2014c-ASR	0.14	0.26	0.13	0.14
ES2014d-ASR	0.04	0.19	0.10	0.08
IS1009a-ASR	0.31	0.47	0.33	0.29
IS1009b-ASR	0.17	0.18	0.20	0.19
IS1009c-ASR	0.19	0.23	0.15	0.14
IS1009d-ASR	0.15	0.30	0.16	0.15
TS3003a-ASR	0.25	0.39	0.28	0.27
TS3003b-ASR	0.22	0.19	0.23	0.21
TS3003c-ASR	0.24	0.24	0.21	0.22
TS3003d-ASR	0.17	0.22	0.18	0.12
TS3007a-ASR	0.22	0.35	0.27	0.20
TS3007b-ASR	0.17	0.18	0.14	0.14
TS3007c-ASR	0.14	0.17	0.16	0.13
TS3007d-ASR	0.15	0.21	0.15	0.16
AVERAGE	**0.19**	**0.27**	**0.19**	**0.18**

Table 5.2: Unsupervised Systems, AMI Corpus, Weighted F-Scores on Manual and ASR using **tf.idf**

MMR=maximal marginal relevance, **CENTR**=LSA centroid, **LSA-SJ**=Steinberger/Ježek SVD, **LSA-Murr**=novel SVD method

Meet	MMR	CENTR	LSA-SJ	LSA-Murr
ES2004a	0.39	0.38	0.38	0.33
ES2004b	0.18	0.18	0.22	0.13
ES2004c	0.19	0.22	0.28	0.24
ES2004d	0.16	0.22	0.24	0.25
ES2014a	0.43	0.55	0.47	0.44
ES2014b	0.18	0.20	0.24	0.24
ES2014c	0.18	0.24	0.28	0.26
ES2014d	0.16	0.20	0.20	0.27
IS1009a	0.44	0.54	0.58	0.54
IS1009b	0.16	0.15	0.16	0.21
IS1009c	0.19	0.28	0.28	0.26
IS1009d	0.29	0.27	0.32	0.33
TS3003a	0.38	0.33	0.34	0.28
TS3003b	0.23	0.25	0.28	0.29
TS3003c	0.31	0.30	0.34	0.30
TS3003d	0.22	0.22	0.23	0.24
TS3007a	0.30	0.23	0.27	0.22
TS3007b	0.18	0.17	0.20	0.20
TS3007c	0.17	0.20	0.27	0.19
TS3007d	0.21	0.17	0.21	0.22
AVERAGE	**0.25**	**0.26**	**0.29**	**0.27**
ES2004a-ASR	0.36	0.39	0.35	0.35
ES2004b-ASR	0.15	0.23	0.20	0.24
ES2004c-ASR	0.20	0.22	0.25	0.22
ES2004d-ASR	0.24	0.26	0.26	0.20
ES2014a-ASR	0.48	0.59	0.50	0.52
ES2014b-ASR	0.20	0.21	0.20	0.18
ES2014c-ASR	0.20	0.23	0.27	0.28
ES2014d-ASR	0.17	0.21	0.18	0.21
IS1009a-ASR	0.51	0.51	0.54	0.54
IS1009b-ASR	0.15	0.17	0.18	0.18
IS1009c-ASR	0.20	0.24	0.22	0.23
IS1009d-ASR	0.34	0.26	0.21	0.26
TS3003a-ASR	0.39	0.40	0.36	0.35
TS3003b-ASR	0.19	0.26	0.27	0.26
TS3003c-ASR	0.31	0.31	0.34	0.31
TS3003d-ASR	0.23	0.27	0.26	0.26
TS3007a-ASR	0.31	0.37	0.30	0.26
TS3007b-ASR	0.16	0.18	0.22	0.18
TS3007c-ASR	0.19	0.17	0.23	0.18
TS3007d-ASR	0.18	0.22	0.19	0.21
AVERAGE	**0.26**	**0.28**	**0.28**	**0.27**

Table 5.3: Unsupervised Systems, AMI Corpus, Weighted F-Scores on Manual and ASR using **su.idf**

MMR=maximal marginal relevance, **CENTR**=LSA centroid, **LSA-SJ**=Steinberger/Ježek SVD, **LSA-Murr**=novel SVD method

Meet	MMR	CENTR	LSA-SJ	LSA-Murr
Bed004	0.11	0.13	0.09	0.09
Bed009	0.11	0.12	0.06	0.07
Bed016	0.16	0.18	0.10	0.15
Bmr005	0.04	0.09	0.05	0.04
Bmr019	0.07	0.10	0.08	0.05
Bro018	0.07	0.15	0.10	0.09
AVERAGE	**0.09**	**0.13**	**0.08**	**0.08**
Bed004-ASR	0.14	0.23	0.12	0.13
Bed009-ASR	0.11	0.13	0.08	0.09
Bed016-ASR	0.22	0.18	0.13	0.18
Bmr005-ASR	0.04	0.07	0.06	0.05
Bmr019-ASR	0.08	0.12	0.08	0.09
Bro018-ASR	0.14	0.17	0.08	0.08
AVERAGE	**0.12**	**0.15**	**0.09**	**0.1**

Table 5.4: Unsupervised Systems, ICSI Corpus, Weighted F-Scores on Manual and ASR Transcripts using **tf.idf**

MMR=maximal marginal relevance, **CENTR**=LSA centroid, **LSA-SJ**=Steinberger/Ježek SVD, **LSA-Murr**=novel SVD method

the weighted f-scores for each meeting using both manual and ASR transcripts.

Using *su.idf* as the term-weighting metric, the overall weighted f-scores for all summarization approaches are again markedly higher than for *tf.idf*. On manual transcripts, the Steinberger/Ježek method is the superior approach with a weighted f-score of 0.16 on average and is significantly better than the the LSA centroid method and novel LSA method ($p<0.1$ and $p<0.05$, respectively), while on ASR that summarization approach suffers considerably and the LSA centroid method is superior with a weighted f-score of 0.13 on average. On ASR, the centroid method is significantly better than the Steinberger/Ježek method.

5.5.3 Discussion

The results of the unsupervised approaches reinforce the findings of Chapter 4 (page 38) that *su.idf* is superior to *tf.idf* as a term-weighting scheme for the purposes of speech summarization on this data. To compare between summarization approaches, the LSA centroid approach is the superior method on both corpora. The summarization approaches as a whole are resistant to ASR errors and generally do not suffer declines in weighted f-score results.

The LSA Centroid method tends to perform similarly with both *su.idf* and *tf.idf*, while the other three unsupervised methods show dramatic increases in f-scores using

Meet	MMR	CENTR	LSA-SJ	LSA-Murr
Bed004	0.13	0.10	0.18	0.10
Bed009	0.13	0.14	0.16	0.09
Bed016	0.14	0.20	0.21	0.18
Bmr005	0.08	0.06	0.06	0.04
Bmr019	0.08	0.07	0.15	0.08
Bro018	0.19	0.18	0.19	0.16
AVERAGE	**0.13**	**0.13**	**0.16**	**0.11**
Bed004-ASR	0.10	0.13	0.10	0.08
Bed009-ASR	0.12	0.11	0.10	0.05
Bed016-ASR	0.18	0.21	0.12	0.12
Bmr005-ASR	0.06	0.07	0.05	0.09
Bmr019-ASR	0.08	0.13	0.05	0.04
Bro018-ASR	0.19	0.16	0.14	0.19
AVERAGE	**0.12**	**0.13**	**0.09**	**0.1**

Table 5.5: Unsupervised Systems, ICSI Corpus, Weighted F-Scores on Manual and ASR Transcripts using **su.idf**

MMR=maximal marginal relevance, **CENTR**=LSA centroid, **LSA-SJ**=Steinberger/Ježek SVD, **LSA-Murr**=novel SVD method

su.idf. One explanation may be that the centroid itself does not actually use the term-weights except in determining the top 20 keywords. This suggests that *tf.idf* rankings may be more reliable than the actual *tf.idf* term-weights.

In general, there are no large differences between the Steinberger/Ježek SVD method and the novel SVD method. They are comparable when applied to the AMI data, and while the former approach is superior on the ICSI manual transcripts, its scores decrease considerably on the ICSI ASR data and are worse on average than the novel SVD method.

Regarding the effect of the different approaches on actual summary output, MMR tends to extract longer units due to its general informativeness score being the sum of dialogue act term scores. While short sentences will be extracted if they contain very high-scoring words, and long sentences will not be extracted if they contain very low-scoring words, there is nonetheless a tendency to extract long dialogue acts on average. This is in contrast with the two SVD approaches, where the summarizers show less favour towards long dialogue acts. The following is a summary excerpt for AMI meeting TS3003c using MMR, illustrating the highest-scoring dialogue acts according to this method:

> **Speaker D:** And on top of that the LCD screen would um help in making the remote control easier to use.
>
> **Speaker B:** We've got um the buttons we have to use. The on-off , sound

on-off , sound higher or lower, um the numbers, uh zero to uh uh nine. Um the general buttons m more general b one button for shifting up and shifting down uh channel.

Speaker D: But if we would make um a changing channels and changing volume button on both sides, that would certainly yield great options for the design of the remote.

Speaker A: Uh requirements are uh teletext, docking station, audio signal, small screen, with some extras that uh button information.

Speaker D: So they would prefer uh a design where the remote control just lies flat in the docking station.

In contrast, these are the top-scoring dialogue acts using the novel SVD method:

Speaker D: So they would prefer uh a design where the remote control just lies flat in the docking station.

Speaker D: Um well the trend-watchers I consulted advised that it b should be, the remote control and the docking station should be telephone-shaped.

Speaker D: So you could imagine that uh the remote control will be standing up straight in the docking station.

Speaker D: Uh the remote control and the docking station should uh blend in in the in the room.

Speaker D: And on top of that the LCD screen would um help in making the remote control easier to use.

Speaker C: Um well the kinetic energy source is rather fancy.

Note that in this particular example, the MMR excerpt shows substantially less redundancy than the SVD method. Five of the six dialogue acts selected by the latter system contain the phrases “remote control” and “docking station” together. While both systems aim to reduce redundancy, the redundancy penalty is dealt with much more explicitly in MMR. With the SVD method, the number of dialogue acts taken from a given topic is determined by the relevant singular value, and so a degree of redundancy is tolerated.

5.6 Results - Feature-Based Approach

This section presents the results of the machine learning approach using a multi-modal features database for both the AMI and ICSI corpora.

5.6.1 AMI Results

For the feature-based approaches, feature subset selection is carried out using a method based on the f statistic as described in Chapter 3 Section 3.5.2 (page 31). The f statistic for each feature is first calculated, and then feature subsets of size n are tried, where n equals 17, 15, 13, 11, 9, 7, 5, and 3, with the n best features included at each step based on the f statistic. The feature subset size with the highest balanced accuracy during cross-validation is selected as the feature set for training. The logistic regression model is then trained on the training data using that subset.

For the AMI data using manual transcripts, the best feature subset according to balanced accuracy is the entirety of the original 17 features. The best five features in order are dialogue act word count, *su.idf* score, dialogue act duration, uninterrupted length of the dialogue act, and *tf.idf* score.

On ASR transcripts, the best feature subset according to balanced accuracy is again the entirety of the 17 features. The best features in order are dialogue act word count, dialogue act uninterrupted length, *su.idf* score, *tf.idf* score, and maximum energy.

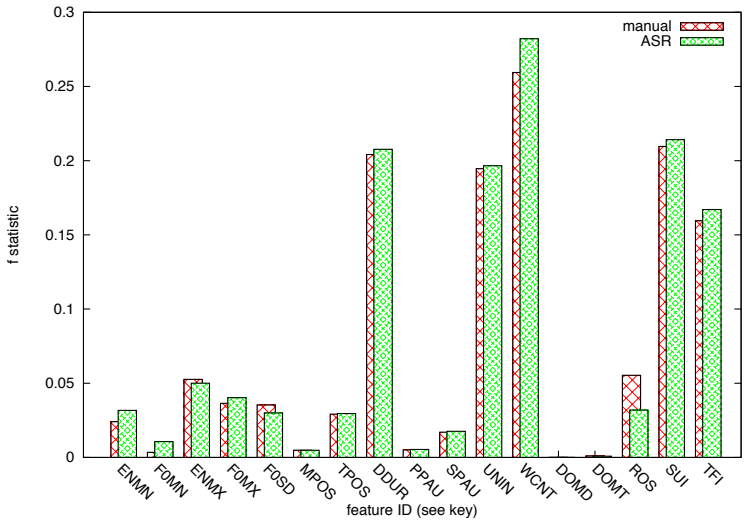


Figure 5.1: F statistics for AMI database features (feature ID key on p. 71)

For the AMI data on using manual transcripts, extractive dialogue acts can be best characterized as having a slightly higher average energy level, a slightly higher average pitch, higher maximum energy and pitch levels, and a higher standard deviation of pitch. They tend to occur slightly earlier in the meetings on average, and later in a speaker's turn. Specifically, the average extractive dialogue act occurs in the third or fourth dialogue act of a turn. The average duration of an extractive dialogue act is 4.45

seconds, compared with just 1.8 seconds for a non-extractive dialogue act. An extractive dialogue act often has a precedent pause, with an average of 0.3 seconds, whereas a non-extractive dialogue act has a negative value of precedent pause, meaning there tends to be overlap between multiple dialogue acts. For subsequent pause, this is reversed, in that extractive dialogue acts have a negative value, signalling speaker overlap, whereas non-extractive dialogue acts exhibit a positive value for subsequent pause. The difference between non-extractive and extractive dialogue acts is even greater in terms of uninterrupted duration of the dialogue act than for total duration of the dialogue act. Extractive dialogue acts differ from non-extractive dialogue acts greatly in terms of word count, with extractive ones average nearly 13 words and non-extractive ones less than 5. There is a small difference between the two classes in terms of speaker dominance, though extractive dialogue acts are slightly more likely to have been uttered by a participant who is more active and dominant in the meeting in general. Extractive dialogue acts have a much higher rate-of-speech than non-extractive, and much higher term-weight scores than non-extractive dialogue acts.

For ASR transcripts, the trends are very similar but with some slight differences. For example, the *su.idf* and *tf.idf* scores on ASR are lower for both classes on average. The rate-of-speech for non-extractive dialogue acts is higher for ASR than for manual transcripts, as the feature is roughly calculated at the word level and the automatic transcript suffers from word insertions. The features relating to word energy and F0 differ slightly because of different word segmentation, but the class differences remain similar.

Figure 5.1 shows the histograms of the feature f statistics using both the manual and ASR databases.

A receiver operator characteristic (ROC) curve plots the ratio of true-positives to false-positives in the classified test data. The ROC curve is a effective evaluation of a classifier because it is not dependant on a particular posterior probability threshold or, in our case, a particular summary length. Figure 5.2 shows the ROC curves for the logistic regression classifiers applied to the AMI test data, using both manual and ASR transcripts. The areas under the curve (AUROC), calculating by divided the area of the graph under the curves into trapezoidal spaces and calculating their individuals areas, are 0.855 for manual transcripts and 0.85 for ASR transcripts. Chance level classification would be 0.5, represented as a diagonal curve from the lower-left to upper-right of the plot.

Table 5.6 lists the weighted f-scores for the 700-word summaries on manual and

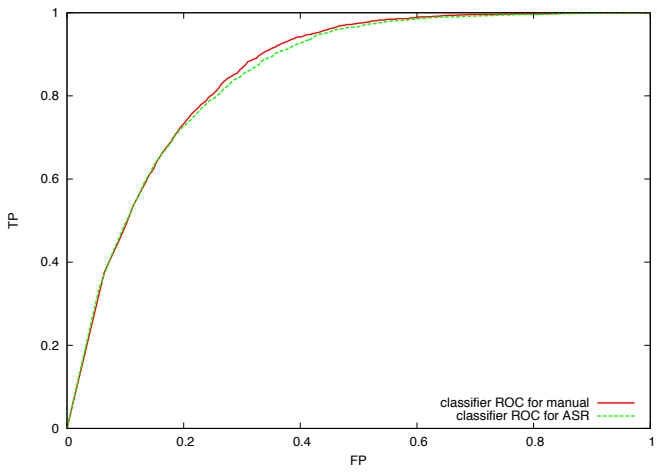


Figure 5.2: ROC Curves for logistic regression Classifiers on AMI data

ASR transcripts using the feature-based approach. There is no significant difference between the manual and ASR f-scores according to paired t-test, and the ASR scores are on average slightly higher.

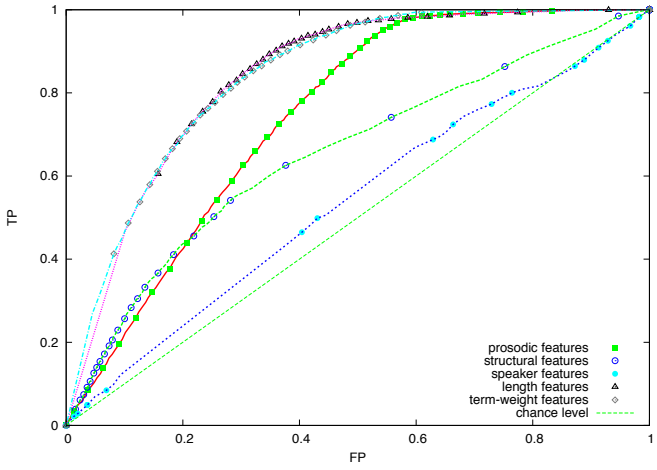
5.6.1.1 Features Analysis

Section 5.6.1 reported a brief features analysis according to each feature's f statistic for the extractive/non-extractive classes. This section expands upon that by examining how useful different subsets of features are for classification on their own. While we found that the optimal subset according to automatic feature subset selection is the entirety of the features database, it is still interesting to examine performance using only certain classes of features on this data. We therefore divide the features into five categories:

- **Prosodic features**: The features of energy, pitch, pause, and rate-of-speech, for a total of 8 features.
- **Length features**: The features of total dialogue act length, uninterrupted length, and dialogue act duration, for a total of 3 features.
- **Speaker features**: The two features of speaker dominance are considered as a class of their own.
- **Structural features**: There are two structural features: the position of the dialogue act in the meeting and the position in the speaker's turn.

Meet	Manual	ASR
ES2004a	0.38	0.40
ES2004b	0.16	0.16
ES2004c	0.17	0.18
ES2004d	0.25	0.24
ES2014a	0.37	0.37
ES2014b	0.16	0.17
ES2014c	0.17	0.18
ES2014d	0.12	0.13
IS1009a	0.38	0.39
IS1009b	0.11	0.11
IS1009c	0.15	0.11
IS1009d	0.22	0.26
TS3003a	0.39	0.39
TS3003b	0.15	0.15
TS3003c	0.15	0.17
TS3003d	0.18	0.19
TS3007a	0.34	0.32
TS3007b	0.15	0.14
TS3007c	0.12	0.12
TS3007d	0.16	0.17
AVERAGE	**0.21**	**0.22**

Table 5.6: AMI Corpus, Weighted F-Scores on Manual and ASR Transcripts for Feature-Based Approach



Fea. Subset	AUROC
Prosodic	0.743
Structural	0.655
Speaker	0.537
Length	0.839
Term-Weight	0.842

Table 5.7: AUROC Values, AMI Corpus, Manual Transcripts

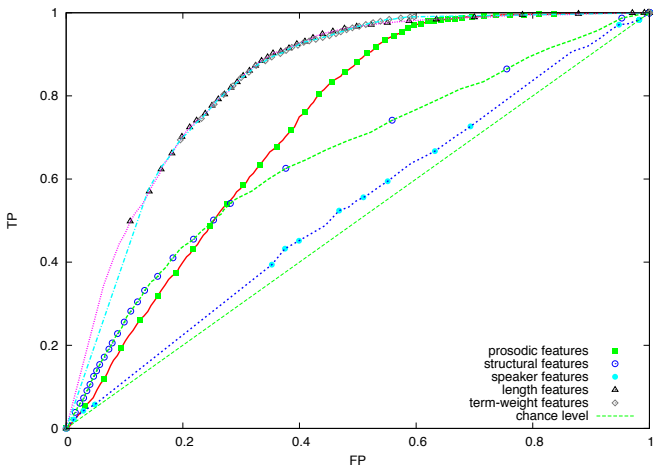
- **Term-weight features**: There are two term-weight features, *tf.idf* and *su.idf*.

One note of interest is that dialogue act duration is not considered as a prosodic feature here. Previous work (Murray et al., 2006; Maskey & Hirschberg, 2005) has reported the duration of the extraction unit as being the best prosodic feature, but as the raw duration is simply correlated to word count, which is also a known useful feature in text summarization, we choose to differentiate between purely prosodic features on the one hand and what we term "length features" on the other. That allows us to examine how "real" prosodic features such as pitch and energy aid summarization classification.

It should also be noted that the "speaker features" are not at all features of the individual dialogue acts, but rather of the speakers themselves. They would not be expected to perform well on their own in a classification task, but are included here for completeness.

Each feature subset is used to train a logistic regression classifier and each classifier is run on the AMI test set, first on manual transcripts then on ASR. We again evaluate the goodness of the classifier using the ROC curve and the AUROC. Figure 5.7 shows the performance of each feature subset classifier relative to chance performance. The AUROCs are as follows: 0.537 for speaker features, 0.655 for structural features, 0.743 for the prosodic features, 0.839 for length features and 0.842 for term-weight features.

The first result to note is that no feature subset classifier AUROC is as good as the AUROC for the full feature set reported in Section 5.6.1, 0.855. The best feature subsets overall are the features of length and of term-weights. The most interesting result, however, is that prosodic features of pitch, energy, pause and rate-of-speech alone



Fea. Subset	AUROC
Prosodic	0.728
Structural	0.655
Speaker	0.532
Length	0.841
Term-Weight	0.835

Table 5.8: AUROC Values, AMI Corpus, ASR Transcripts

result in very respectable classification. It is encouraging and worth emphasizing that prosodic features other than durational features are very useful for extractive classification. It is also slightly surprising that the two structural features alone performed as well as they did, well above chance levels using only the dialogue act position in the meeting and in the speaker turn.

Figure 5.8 shows the ROC curves for the classifiers applied to the ASR database. The AUROCs are as follows: 0.532 for the speaker features, 0.655 for the structural features, 0.728 for the prosodic features, 0.841 for the length features, and 0.835 for the term-weight features. The trends are much the same as with manual transcripts, but with a few intriguing differences. Prosodic features still perform very well but slightly lower than on manual transcripts. This result may seem counter-intuitive at first. With an errorful ASR transcript, it might be expected that prosodic features would be more valuable and term-weight features less valuable. Of course, the prosodic features rely on word segmentation and the prosodic data can become noisy when word boundaries are incorrect. The term-weight features are also slightly worse, while length features are slightly more effective on the ASR data.

5.6.2 ICSI Results

For the ICSI corpus using manual transcripts, the optimal feature subset consists of 15 features according to balanced accuracy, excluding mean F0 and precedent pause. The best 5 features according to the f statistic are dialogue act word count, uninterrupted length, *su.idf* score, *tf.idf* score and dialogue act duration. Figure 5.3 shows the

histograms for the feature f statistics using both the manual and ASR databases.

The optimal subset for ASR transcripts was again 15 features, excluding mean F0 and precedent pause, with the best 5 features being dialogue act word count, uninterrupted length, *su.idf*, dialogue act duration, and *tf.idf*.

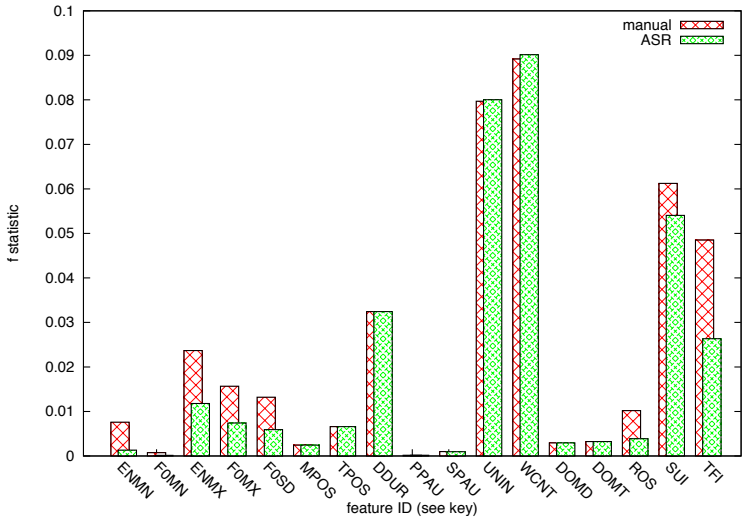


Figure 5.3: F statistics for ICSI database features

The extractive dialogue acts in the ICSI corpus using manual transcripts can be characterized as having high average energy and pitch levels, high maximum pitch and energy levels, and a high pitch standard deviation. They tend to occur earlier in the meeting on average, and later in a speaker's turn. The average duration is more than 4.5 seconds, compared with less than 2 seconds for the average non-extractive dialogue act. Similar to the AMI corpus, there tends to be a long precedent pause, but a negative subsequent pause, i.e. overlap at the end of the dialogue act. The uninterrupted duration of the dialogue act is much greater than for non-extractive dialogue acts. The average word count is more than 15, compared with just over 6 for the negative class. Extract-worthy dialogue acts tend to be spoken by meeting participants who are more dominant in the meeting as a whole. The rate-of-speech is considerably higher for the positive class, and both term-weight scores are much higher for extractive dialogue acts than for non-extractive dialogue acts.

Figure 5.4 shows the ROC curves for the logistic regression classifiers applied to the ICSI data for both manual and ASR transcripts. The AUROC for manual transcripts is 0.818 and for ASR transcripts it is 0.824.

Table 5.9 shows the weighted f-scores for the 700-word summaries for both manual

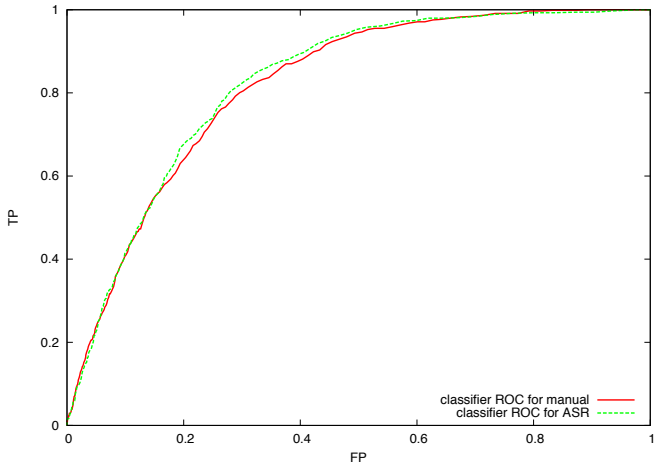


Figure 5.4: ROC Curves for logistic regression Classifiers on ICSI data

Meet	Manual	ASR
Bed004	0.13	0.13
Bed009	0.17	0.17
Bed016	0.13	0.21
Bmr005	0.14	0.12
Bmr019	0.10	0.11
Bro018	0.12	0.15
AVERAGE	**0.13**	**0.15**

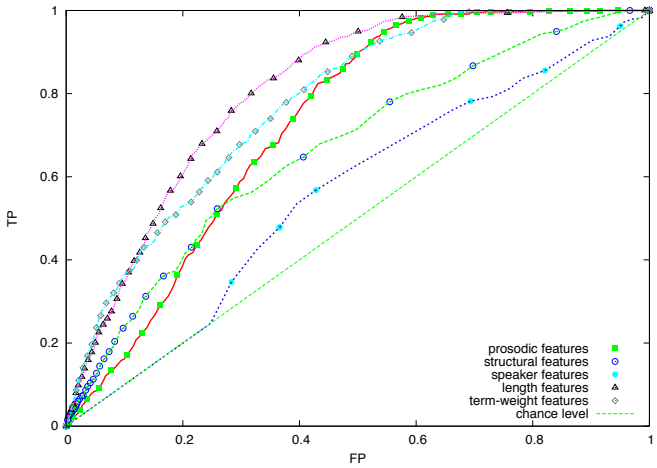
Table 5.9: ICSI Corpus, Weighted F-Scores on Manual and ASR Transcripts for Feature-Based Approach

and ASR transcripts using the feature-based approach. As with the AMI corpus, there is no significant difference between manual and ASR results and the ASR average is slightly higher.

5.6.2.1 Features Analysis

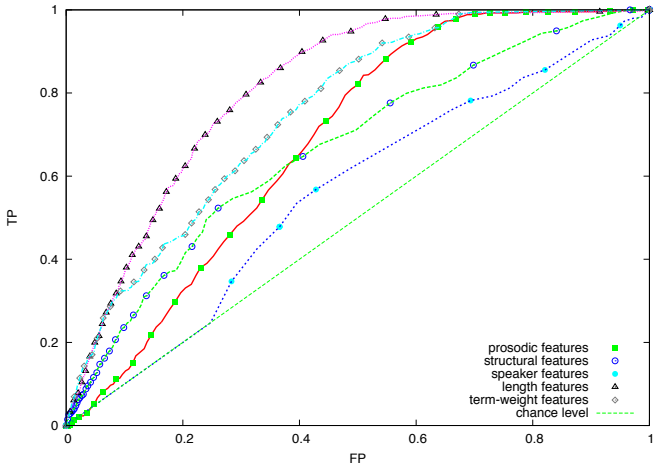
In this section we report the result of the separate feature subsets on classification. The five subsets are the same as reported above for the AMI data: prosodic features, structural features, speaker features, length features, and term-weight features.

The ROC curves for each classifier applied to manual transcripts are shown in Figure 5.10. The AUROCs for the relevant feature subsets are as follows: 0.559 for speaker features, 0.668 for structural features, 0.728 for prosodic features, 0.776 for term-weight features, and 0.809 for length features. Again, no subset alone is superior to using all of the features for classification, though for manual transcripts the length



Fea. Subset	AUROC
Prosodic	0.728
Structural	0.668
Speaker	0.559
Length	0.809
Term-Weight	0.776

Table 5.10: AUROC Values, ICSI Corpus, Manual Transcripts



Fea. Subset	AUROC
Prosodic	0.683
Structural	0.667
Speaker	0.559
Length	0.812
Term-Weight	0.755

Table 5.11: AUROC Values, ICSI Corpus, ASR Transcripts

features subset is competitive. An interesting result is that for the ICSI data the length feature is considerably superior to the term-weight features. We also find, similar to the result with AMI data, that prosodic features alone are able to perform respectable summarization classification.

The ROC curves for each classifier applied to ASR transcripts are shown in Figure 5.11. The AUROCs for the relevant feature subsets are as follows: 0.559 for speaker features, 0.667 for structural features, 0.683 for prosodic features, 0.755 for term-weight features, and 0.812 for length features. The AUROC for prosodic features is noticeably worse when applied to ASR transcripts. For the ICSI corpus, length features are much more useful than term-weight features, whereas for the AMI corpus those two feature subsets were more comparable.

5.6.3 Combined Training Data

As an attempt to investigate domain-independent features for meeting summarization, the training data for the AMI and ICSI corpora are combined to create a single training set. The test sets for both corpora are then classified using the combined model. Feature subset selection is carried out as described previously (e.g. Chapter 3 Section 3.5.2, page 31), and in this case the optimal subset according to balanced accuracy is n=15 for manual transcripts, with mean F0 and precedent pause excluded. The best five features in order according to the f statistic are *su.idf* score, dialogue act word count, uninterrupted length, *tf.idf* score, and dialogue act duration. For ASR transcripts, the optimal subset consists of the entirety of the 17 features. The best five features according to the f statistic are word count, *su.idf* score, uninterrupted length, *tf.idf* score and duration.

It can be seen that there is no increase in summarization performance after combining the training data. Table 5.12 shows the AMI results for 700 word summaries, with human transcript summaries scoring slightly higher and ASR transcript summaries scoring the same. Table 5.13 shows the results for the ICSI corpus, with the manual and ASR transcript summaries both scoring slightly lower on average. The AUROCs are lower for all four classifiers. Both AMI classifiers have AUROCs of 0.83 while the ICSI classifiers have AUROCs of 0.78 and 0.81 for manual and ASR, respectively.

Meet	Manual	ASR
ES2004a	0.34	0.39
ES2004b	0.16	0.16
ES2004c	0.17	0.13
ES2004d	0.18	0.18
ES2014a	0.41	0.35
ES2014b	0.16	0.16
ES2014c	0.16	0.15
ES2014d	0.12	0.15
IS1009a	0.41	0.45
IS1009b	0.15	0.10
IS1009c	0.18	0.11
IS1009d	0.32	0.28
TS3003a	0.36	0.36
TS3003b	0.20	0.17
TS3003c	0.27	0.12
TS3003d	0.17	0.21
TS3007a	0.23	0.29
TS3007b	0.17	0.15
TS3007c	0.18	0.12
TS3007d	0.16	0.20
AVERAGE	**0.23**	**0.22**

Table 5.12: AMI Corpus, Weighted F-Scores on Manual and ASR Transcripts for Feature-Based Approach, Combined Training Data

Meet	Manual	ASR
Bed004	0.12	0.13
Bed009	0.14	0.15
Bed016	0.13	0.21
Bmr005	0.12	0.12
Bmr019	0.09	0.10
Bro018	0.14	0.15
AVERAGE	**0.12**	**0.14**

Table 5.13: ICSI Corpus, Weighted F-Scores on Manual Transcripts for Feature-Based Approach, Combined Training Data

5.6.4 Discussion

For the feature-based machine-learning approaches, we find that optimal results are derived by using a variety of multi-modal features for this data, including lexical, prosodic, structural, length and speaker features. For both the AMI and ICSI corpora, the optimal feature subsets include each of these feature types. This is attested both through feature subset selection on the training data, where a wide variety of features result in superior balanced accuracy during cross-validation, and in classification of the test data, where the best AUROC results are derived by using a combination of multi-modal features. And though term-weight and length features can at times perform very well, the only consistent set is the entire feature set. For example, term-weight features alone result in respectable classification on the AMI test set using manual transcripts, but less well on the AMI ASR test set, and much less well on the ICSI data in general. The length features are more consistent but never superior to the full feature set. The disadvantage of relying on length features is that fewer dialogue acts are extracted as a result of favoring very long dialogue acts, thereby lowering recall scores. As seen in this excerpt of AMI meeting TS3003c summarized using only length features, three dialogue acts alone account for about 120 words of the summary.

> **Speaker D:** 'Cause we would have to make one w uh control which would fit in with a wooden cover and a plastic cover. The more original one, or the more standard one.
>
> **Speaker B:** We've got um the buttons we have to use. The on-off, sound on-off, sound higher or lower, um the numbers, uh zero to uh uh nine. Um the general buttons m more general b one button for shifting up and shifting down uh channel.
>
> **Speaker B:** Um double push push um, if double click, um so uh you get uh big uh subtitles, for uh people uh um uh which c f uh who can't uh read small uh subtitles .

Compression techniques can help distill these dialogue acts to their essence, but in Chapter 8 Section 8.3 (page 165) we also consider using extraction units of a finer granularity than entire dialogue acts.

To contrast, summaries created using only prosodic features such as pitch, energy and rate-of-speech do not favour longer dialogue acts at all and as a result have higher recall scores. Here we provide an excerpt of meeting TS3003c generated using only these prosodic features:

> **Speaker A:** Look I've got a new remote control, and uh
>
> **Speaker C:** Because, like on your mobile phone, it's always above.

Speaker A: I think uh elderly people just like to have everything in place.

Speaker D: I think that was a very good point.

Speaker A: Uh f I think first of all we have to see uh it is possible to introduce kinetic energy in our budget, I think.

Speaker C: About the components design.

Speaker D: if you'd allow me to go to the flat board, SMARTboard.

To compare the supervised and unsupervised methods presented in this chapter, it is worth pointing out that the use of weighted f-score as the overall metric obscures one fact. While the best unsupervised methods from Section 5.5 (page 71) have comparable f-scores to the supervised methods described in this section, their precision is much lower. For example, the average weighted precision for the machine-learning approach on the AMI test set with manual transcripts is 0.64, whereas the LSA Centroid method that performed best overall among the unsupervised methods has an average weighted precision of only 0.54. In terms of weighted precision, the feature-based approach is significantly better ($p<0.05$). The difference is even more striking on the ICSI data. The machine-learning approach and the LSA centroid method applied to manual transcripts have comparable f-scores, but the weighted precision for the former is 0.46 compared with 0.28 for the latter (again significant at $p<0.05$). The unsupervised approaches have comparable or even higher f-scores because their recall scores are substantially higher and precision is lower. The reason for the supervised approach having lower recall is that the length features are very indicative of informativeness, and so the units of extraction tend to be very long in the machine-learning approach. In contrast, the unsupervised methods will sometimes extract shorter units and therefore extract more dialogue acts for a given compression rate. In Chapter 8 Section 8.3 (page 165), we explore the use of spurts instead of dialogue acts as our unit of extraction, for the dual purposes of faster segmentation of the speech stream and a finer level of granularity for our extractive summarization units.

This difference between precision and recall also relates to the comparison of our best machine-learning results to human extraction performance. The creation of human extractive summaries at the same compression level is described in detail in Chapter 4 Section 4.1.3 (page 46). If we compare average weighed f-scores, the human summarizers are considerably better for both corpora. The reason again is that more units will be extracted for a given compression rate because some of the units are shorter, whereas our logistic regression model favors longer extraction units. However, if we compare solely in terms of weighted precision, we see that we attain human-level pre-

Meet	A-P	A-R	A-F	H-P	H-R	H-F
ES2004a	0.60	0.27	0.38	0.67	0.56	0.61
ES2004b	0.68	0.09	0.16	0.83	0.25	0.39
ES2004c	0.70	0.10	0.17	0.58	0.18	0.28
ES2004d	1.09	0.14	0.25	1.03	0.30	0.45
ES2014a	0.63	0.26	0.37	0.82	0.58	0.68
ES2014b	0.78	0.09	0.16	0.80	0.22	0.35
ES2014c	0.85	0.10	0.17	1.21	0.32	0.51
ES2014d	0.46	0.07	0.12	0.63	0.25	0.36
IS1009a	0.61	0.28	0.38	1.16	0.70	0.86
IS1009b	0.57	0.06	0.11	1.15	0.25	0.41
IS1009c	0.40	0.09	0.15	0.72	0.33	0.45
IS1009d	0.64	0.13	0.22	1.10	0.44	0.63
TS3003a	0.56	0.30	0.39	0.68	0.47	0.52
TS3003b	0.56	0.09	0.15	0.98	0.28	0.44
TS3003c	0.57	0.09	0.15	0.93	0.30	0.45
TS3003d	0.49	0.11	0.18	0.70	0.30	0.42
TS3007a	0.60	0.23	0.34	0.86	0.51	0.63
TS3007b	0.61	0.08	0.15	0.65	0.19	0.29
TS3007c	0.56	0.07	0.12	0.92	0.24	0.38
TS3007d	0.75	0.09	0.16	0.86	0.21	0.34
AVERAGE	**0.64**	**0.14**	**0.21**	**0.87**	**0.35**	**0.47**

Table 5.14: AMI Human Summarization Scores Comparison

A-P=automatic summarizer precision, **A-R**=automatic summarizer recall, **A-F**=automatic summarizer f-score, **H-P**=human summarizer precision, **H-R**=human summarizer recall, **H-F**=human summarizer f-score

cision on the ICSI corpus as a whole, and human-level performance on several AMI test set meetings. Tables 5.14 and 5.15 compare the best automatic classification results with human classification results for the AMI and ICSI corpora, respectively, using the ASR-aligned databases. For the ICSI corpus, the results on manual transcripts are actually superior to human performance according to weighted precision, averaging 0.46 compared with 0.41 for humans.

The ICSI scores overall are lower than the AMI scores, reflecting lower inter-annotator agreement on that corpus as reported in Chapter 3 Section 3.3.2.1 (page 30). And whereas we achieve human-level performance on the ICSI corpus, there is still a substantial gap between machine and human performance on the AMI corpus.

5.7 Conclusions

In this chapter we have presented a variety of supervised and unsupervised extractive summarization techniques for spontaneous meeting speech. Amongst our unsuper-

Meet	A-P	A-R	A-F	H-P	H-R	H-F
Bed004	0.33	0.08	0.13	0.41	0.17	0.23
Bed009	0.48	0.07	0.12	0.39	0.17	0.23
Bed016	0.40	0.08	0.14	0.42	0.14	0.20
Bmr005	0.70	0.04	0.08	0.52	0.12	0.19
Bmr019	0.43	0.06	0.10	0.40	0.14	0.21
Bro018	0.41	0.09	0.14	0.34	0.12	0.16
AVERAGE	**0.46**	**0.07**	**0.12**	**0.41**	**0.14**	**0.20**

Table 5.15: ICSI Human Summarization Scores Comparison

A-P=automatic summarizer precision, **A-R**=automatic summarizer recall, **A-F**=automatic summarizer f-score, **H-P**=human summarizer precision, **H-R**=human summarizer recall, **H-F**=human summarizer f-score

vised approaches, we find that the novel LSA centroid method consistently performs the best. More broadly, we find that all of the unsupervised approaches are made much more effective by using a term-weighting scheme more robust than the standard *tf.idf* scheme. Specifically, the novel term-weighting method *su.idf* was very useful for both the unsupervised systems and as a feature in the supervised model.

For the supervised method using a logistic regression classifier trained on labelled data, we find that using a variety of features from the data yields optimal performance, superior to simply treating the data as noisy text and using only text summarization methods. Even prosodic features alone yield decent summarization results according to the classifier AUROC measures. This finding is encouraging in that we considered length features to be a separate class from the prosodic features, and still find prosodic features relating to F0, energy, pause and rate-of-speech to be very effective indicators of informativeness.

Weighted f-score results are generally lower on the ICSI corpus than on the AMI corpus, reflecting the lower inter-annotator agreement on that data, and our summarizer performance on that data is actually closer to human-level performance than with the AMI data. Regarding the features analysis, the length features are considerably more useful than term-weight features for creating summaries of the ICSI test set, whereas these two feature subsets are more comparable for the AMI test set.

Chapter 6

Extrinsic Evaluation - A Decision Audit Task

6.1 Introduction

In previous chapters, the automatic summaries were evaluated *intrinsically* by scoring them according to multiple human annotations of informativeness. That is, they were evaluated according to how well their information content matched the information content of gold-standard summaries. A more comprehensive and reliable evaluation of the quality of a given summary, however, is the degree to which it aids a real-world *extrinsic* task: an indication not just of how informative the summary is, but how useful it is in a realistic task. As mentioned in the introduction to this thesis, the purpose of these summaries is not to serve as stand-alone indicators of meeting information content, but to aid user *navigation* of the entire meeting content. The meeting summaries are meant to index the greater overall meeting record. We therefore design an extrinsic task that models a real-world information need, create multiple experimental conditions comprised of various representations of meeting information content, and enlist subjects to participate in the task.

The chosen task is a *decision audit*, wherein a user must review previously held meetings in order to determine how a given decision was reached. This involves the user determining what the final decision was, which alternatives had previously been proposed, and what the arguments for and against the various proposals were. The reason this task was chosen is that it represents one of the key use cases for AMI technologies - that of aiding *corporate memory*, the storage and management of a organization's knowledge, transactions, decisions, and plans. A organization may find

itself in the position of needing to review or explain how it came to a particular position or why it took a certain course of action. When business meetings are archived and summarized, this task should be made much more efficient.

The decision audit represents a complex information need that cannot be satisfied with a simple one-sentence answer. Relevant information will be spread throughout several meetings and may appear at multiple points in a single discussion thread. Because the decision audit does not only involve knowing *what* decision was made but also determining *why* the decision was made, the person conducting the audit will need to understand the evolution of the meeting participants' thinking and the range of factors that led to the ultimate decision. For a particular decision audit task, the decision itself may be a given. Because the person conducting the decision audit does not know which meetings are relevant to the given topic, there is an inherent relevance assessment task built into this overall task. Their time is limited and they cannot hope to scan the meetings in their entirety and so must focus on which meetings and meeting sections seem most promising. It should be noted, however, that none of the summaries described in the conditions below were generated with this particular information need in mind. They are strictly generic.

6.2 Related Extrinsic Evaluation Work

This section describes previous extrinsic evaluations relating either to summarization specifically, or else to the browsing of multi-party interactions more generally. We then describe how our decision audit browsers fit into a typology of multi-media interfaces.

In the field of text summarization, a commonly used extrinsic evaluation has been the *relevance assessment* task (Mani, 2001b). In such a task, a user is presented with a description of a topic or event and then must decide whether a given document (e.g. a summary or a full-text) is relevant to that topic or event. Such schemes have been used for a number of years and on a variety of projects (Jing et al., 1998; Mani et al., 1999; Harman & Over, 2004). Due to problems of low inter-annotator agreement on such ratings, Dorr et al. (2005) proposed a new evaluation scheme that compares the relevance judgement of an annotator given a full text with that same annotator given a condensed text.

Another type of extrinsic evaluation for summarization is the *reading comprehension* task (Hirschman et al., 1999; Morris et al., 1992; Mani, 2001b). In such an evaluation, a user is given either a full source or a summary text and is then given a

multiple-choice test relating to the full source information. A system can then calculate how well they perform on the test given the condition. This evaluation framework relies on the idea that truly informative summaries should be able to act as substitutes for the full source. This doesn't hold true for certain classes of summaries such as query-dependent or indicative summaries. In the case of query-dependent summaries, it would be expected that reading a summary would yield better comprehension than reading the full source if the full source contained a great deal of information irrelevant to the task questions. Of course, the extrinsic task could be tailored to suit various summary types, e.g. by comparing an automatically generated query-dependent summary to a human-authored query-dependent summary for a reading comprehension task.

In the DUC conferences [1], human judges assign a pseudo-extrinsic *responsiveness* score to each machine summary, representing how well the given summary satisfies the information need in the query. This is not a true task-based extrinsic evaluation, but does give a sense of the potential utility of the summary in light of the query. Daumé and Marcu (2005) have suggested that DUC adopt an extrinsic evaluation framework in future years, specifically suggesting a relevance prediction task, and pointing out that some of the considerable time and labor required for annotations such as for the Pyramid scheme could be spent implementing a simple task-based evaluation.

Wellner et al. (2005) introduced the Browser Evaluation Test (BET), in which *observations of interest* are collected for each meeting, e.g. the observation "Susan says the footstool is expensive." Each observation is presented as both a positive and negative statement and the user must decide which statement is correct by browsing the meetings and finding the correct answer. It is clear that such a set-up could be used to evaluate summaries and to compare summaries with other information sources. We chose not to use this evaluation paradigm, however, because the observations of interest tend to be skewed towards a keyword search approach, where it would always be simpler just to search for a word such as "footstool" rather than read a summary. It might be possible to set up the BET in such a manner that the observations of interest are less biased towards a particular type of content extraction, but we instead choose a more complex information need for our evaluation. There are some similarities between the BET and the TREC Interactive Track (Hersh & Over, 2001), as the latter examines the ability of a human searcher to answer a set of questions given a particular information retrieval system. In the Interactive Track, there is a focus not only on the result but on the searching process, an idea that is inherent in the decision audit

[1] http://duc.nist.gov

task as well.

Also on the AMI project, the Task-Based Evaluation (TBE) (Kraaij & Post, 2006) evaluates multiple browser conditions containing various information sources relating to a series of AMI meetings. Participants are brought in four at a time and are told that they are replacing a previous group and must finish that group's work. In essence, the evaluation involves re-running the final meetings of the series with new participants. The participants are given information related to the previous group's initial meetings and must finalize the previous group's decisions as best as possible given what they know. The reason we did not choose the TBE for this summarization evaluation is that the TBE evaluation relies on lengthy post-questionnaire results rather than more objective criteria. For example, users are asked to rate the statement "There is no better information source than this browser," when they may not in fact be in the position to know whether or not there are better options. The TBE is also more costly to run than our decision audit task, as it requires having groups of four people spend an afternoon reviewing previous meetings and conducting their own meetings, which are also recorded, whereas the decision audit is an individual task.

The SCANMail browser (Hirschberg et al., 2001; Whittaker et al., 2002) is an interface for managing and browsing voicemail messages, with multi-media components such as audio, ASR transcripts, audio-based paragraphs, and extracted names and phone numbers. To evaluate the browser and its components, the authors compared the SCANMail browser to a state-of-the-art voicemail system on four key tasks: scanning and searching messages, extracting information from messages, tracking the status of messages (e.g. whether or not a message has been dealt with), and archiving messages. Both in a think-aloud laboratory study and a larger field study, users found the SCANMail system outperformed the comparison system for these extrinsic tasks. The field study in particular yielded several interesting findings. In 24% of the times that users viewed a voicemail transcript with the SCANMail system, they did not resort to playing the audio. This testifies to the fact that the transcript and extracted information can, to some degree, act as substitutes for the signal, which user comments also back up. On occasions when users did play the audio, 57% of the time they did not play the entire audio. Most interestingly, 57% of the audio play operations resulted from clicking within the transcript. The study also found that users were able to understand the transcripts even with recognition errors, partly by having prior context for many of the messages.

The SpeechSkimmer browser (Arons, 1997) is an audio-based browser incorporat-

ing skimming, compression and pause-removal techniques for the efficient navigation of large amounts of audio data. The authors conducted a formative usability study in order to refine the interface and functionality of SpeechSkimmer, recruiting participants to find several pieces of relevant information within a large portion of lecture speech using the browser. Results were gleaned both from a think-aloud experiment structure as well as follow-up questions on ease of use. The researchers found that experiment participants often began the task by listening to the audio at normal speed to first get a feel for the discussion, and subsequently made good use of the skimming and compression features to increase search efficiency.

Whittaker et al. (2008) described a task-oriented evaluation of a browser for navigating meeting interactions. The browser contains a manual transcript, a visualization of speaker activity, audio and video streams with play, pause and stop commands, and artefacts such as slides and whiteboard events (the slides, but not the whiteboard events, are indices into the meeting record). Users were given two sets of questions to answer, the first set consisting of general "gist" question about the meeting, and the second set comprised of questions about specific facts within the meeting. There were 10 questions in total to be answered. User responses were subsequently scored on correctness compared with model answers. There are several interesting findings from this task-based evaluation. While general performance was not high, users found it much easier to answer specific questions than "gist" questions using this browser setup. This has special relevance for our work, as certain types of information needs might be easily satisfied without recourse to derived data such as summaries or topic segments, but getting the general gist of the meeting seems to be much more difficult. Very interestingly, users often felt that they had performed much better than they actually had. Specifically, users seemed to be unaware that they had missed relevant or vital information and felt that they had provided comprehensive answers. Across the board, participants focused on reading the transcript rather than beginning with the audio and video records directly.

6.2.1 Multi-Modal Browser Types

Tucker and Whittaker (2004) provided an overview of the mechanisms available for browsing multi-modal meetings. They established a four-way browser classification: audio-based browsers, video-based browsers, artefact-based browsers, and derived data browsers. With audio-based browsers, the audio recordings of the meeting are

the main focus, and are sometimes coupled with a visual index for navigating through the audio record by clicking on, for example, speaker segments (Kimber et al., 1995). Other audio browsers feature the facility to alter playback speed or to compress the audio in some fashion (Arons, 1997; Tucker & Whittaker, 2006).

With video browsers, both audio and video are provided to the user, but the focus is on the video. These browsers are highly dependent on the actual environment of the meetings, as in some cases each participant will have a camera trained solely on them with additional room-view cameras (Carletta et al., 2005), and in other cases there may be a single panoramic camera for recording the meetings (Lee et al., 2002). As with audio browsers, there may be a separate visual index or a facility for speed-up or compression. Another possibility for video browsers is to extract *keyframes* or video grabs, which are relevant static images from the video stream, and then present the keyframes in a story-board or comics format (Girgensohn et al., 2001; Kleinbauer et al., 2007).

The third class as established by Tucker and Whittaker is comprised of artefact-based browsers, with artefacts being information recorded in the meeting other than the audio/video streams. For the AMI meetings, artefacts include slides, notes, whiteboard drawings, and emails. Each of these can be very informative, and by synchronizing all of these sources of information to the audio/video record, a person using the browser can more fully get a sense of the meeting interactions. Furthermore, artefacts such as slides can be useful for indexing into the audio/video record.

The fourth class is comprised of browsers incorporating derived data forms. These browsers feature components that result from in-depth analysis of the meetings rather than simply recording various phenomena in the meetings. These components include ASR transcripts, topic segmentation, automatically generated summaries, dialogue act segmentation and labelling, and emotion or sentiment detection. These components provide structure and semantics to the meeting record, and again can act as efficient indices into the meeting record.

In light of this classification scheme, our decision audit browsers are video browsers incorporating derived data forms. Although other incarnations of our browsers contain meeting artefacts such as slides, we simplify the browsers as much as possible for this task by putting the focus on derived data forms and their usefulness for browsing the meeting records. Each version of the experimental browser is built using the JFerret (Wellner et al., 2004), an easily modifiable multi-media browser framework[2].

[2]http://www.idiap.ch/mmm/tools/jferret

6.3 Task Setup

The data for the extrinsic evaluation is one meeting series ES2008 from the AMI corpus, comprised of 4 related, sequential meetings. The particular meeting series is chosen because it has been used in previous AMI extrinsic evaluations and the participant group in that series worked well together on the task. The group took the task seriously and exhibited deliberate and careful decision-making processes in each meeting and across the meetings as a whole.

6.3.1 Task Overview

The extrinsic task is an individual task, unlike the AMI TBE, described above, which was a group-based scenario task. We recruited only participants who were native English speakers and who had not participated in previous AMI experiments or data collection. 10 subjects were run per condition, for a total of 50 subjects. For each condition, 6 participants were run in Edinburgh and 4 were run at DFKI, an AMI partner. The experimental setups for the two locations were as identical as possible, with comparable desktop machines running Linux, 17-inch monitors, identical browser interfaces, and the same documents used in each location, as described below. And though DFKI is a German institution, they recruited only native English speakers, primarily from their student and researcher populations.

Each participant is first given a pre-questionnaire relating to background, computer experience and experience in attending meetings (see Appendix A, page 176). In the case that the participant regularly participates in meetings, we ask how they normally prepare for a meeting, e.g. using their own notes, consulting with other participants, etc.

Each participant is then given general task instructions (Appendix A). These instructions explain the meeting browser in terms of the information provided in the browser and the navigation functions of the browser, the specific information need they are meant to satisfy in the task, and a notice of the allotted time for the task. The total time allotted is 45 minutes, which includes both searching for the information and writing up the answer. This amount of time is based on the result of an individual pilot task for Condition EAM, extractive summarization on manual transcripts.

The portion of the instructions detailing the specific task reads as follows:

> We are interested in the group's decision-making ability, and therefore ask you to evaluate and summarize a particular aspect of their discussion.

Condition	Description
KAM	Top 20 keywords
EAM	Extractive summary of manual transcripts
EAA	Extractive summary of ASR transcripts
AMM	Human abstracts
ASM	Semi-Automatic abstracts

Table 6.1: Experimental Conditions

> The group discussed the issue of separating the commonly-used functions of the remote control from the rarely-used functions of the remote control. What was their final decision on this design issue? Please write a short summary (1-2 paragraphs) describing the final decision, any alternatives the participants considered, the reasoning for and against any alternatives (including why each was ultimately rejected), and in which meetings the relevant discussions took place.

This particular information need is chosen because the relevant discussion manifested itself throughout the 4 meetings, and the group went through several possibilities before designing an eventual solution to this portion of the design problem. In the first meeting, the group discussed the possibility of creating two separate remotes. In the second meeting, it was proposed to have simple functions on the remote and more complex functions on a sliding compartment of the remote. In the third meeting, they decided to have an on-screen menu for complex functions, and in the final meeting they finalized all of the details and specified the remote buttons. A participant in the decision audit task therefore would have to consult each meeting to be able to retrieve the full answer to the task's information need.

While in this case the participant must determine the decision that was made *and* the reasons behind the decision, in theory the decision audit could be set up in such a way that the decision itself is a given and only the reasoning behind the decision must be determined.

6.3.2 Experimental Conditions

There are 5 conditions run in total: one baseline condition, two extractive conditions and two abstractive conditions.

The baseline condition, Condition KAM, consists of a browser with manual transcripts, audio/video record, and a list of the top 20 keywords in the meeting. The keywords are determined automatically using *su.idf*, a weighting scheme described in

Chapter 4 (page 38). Figure 6.1 shows a screen-shot for the browser in Condition KAM. Though this is a baseline condition, the fact that it utilizes *manual* transcripts gives users in this condition a possible advantage over users in conditions with ASR. In this respect, it is a challenging baseline.

Conditions EAM and EAA present the user with a transcript, audio/video record and an automatically-generated extractive summary of each meeting, with the difference between the conditions being that the latter is based on ASR and the former on manual transcripts. The features used are the same as in Chapter 5 (page 63), but with support vector machine classifiers instead of logistic regression classifiers. The lengths of the respective extractive summaries are based on the lengths of the manual extracts for each meeting: approximately 1000 words for the first meeting, 1900 words for the second and third meetings, and 2300 words for the final meeting. These lengths correlate to the lengths of the meetings themselves and represent compressions of approximately 40%, 32%, 32% and 30%, respectively. Figure 6.2 shows a screenshot for the browser in Conditions EAM and EAA.

Condition AMM is the gold-standard condition, a human-authored abstractive summary. Each summary is divided into subsections: decisions, actions, goals and problems. These abstractive summaries vary in length. Each abstractive sentence is normally also linked to one or more transcript dialogue acts, making the experimental condition a *hybrid* of abstractive and extractive. Because this is a decision audit task and the abstractive summary provided in this condition has a "decisions" subsection, this is considered to be a challenging gold-standard condition to match. Figure 6.3 shows a screen-shot for the browser in Condition AMM.

Condition ASM presents the user with an semi-automatically generated abstractive summary, described by Kleinbauer et al. (2007). This summarization method utilizes automatic topic segmentation and topic labels, and finds the most commonly mentioned content items in each topic. A sentence is generated for each meeting topic indicating what was discussed, and these sentences are linked to the actual dialogue acts in the discussion. These summaries rely on *manual* transcripts, and so Condition EAA is the only ASR condition in this experiment. The Condition ASM summaries are not fully automatic, as they rely on manual annotation of propositional content. Figure 6.4 shows a screen-shot for the browser in Condition ASM.

Table 6.1 lists and briefly describes the experimental conditions. The three-letter ID for each condition corresponds to **k**eywords/**e**xtracts/**a**bstracts, **a**utomatic/**s**emi-automatic/**m**anual algorithms, and **a**utomatic/**m**anual transcripts.

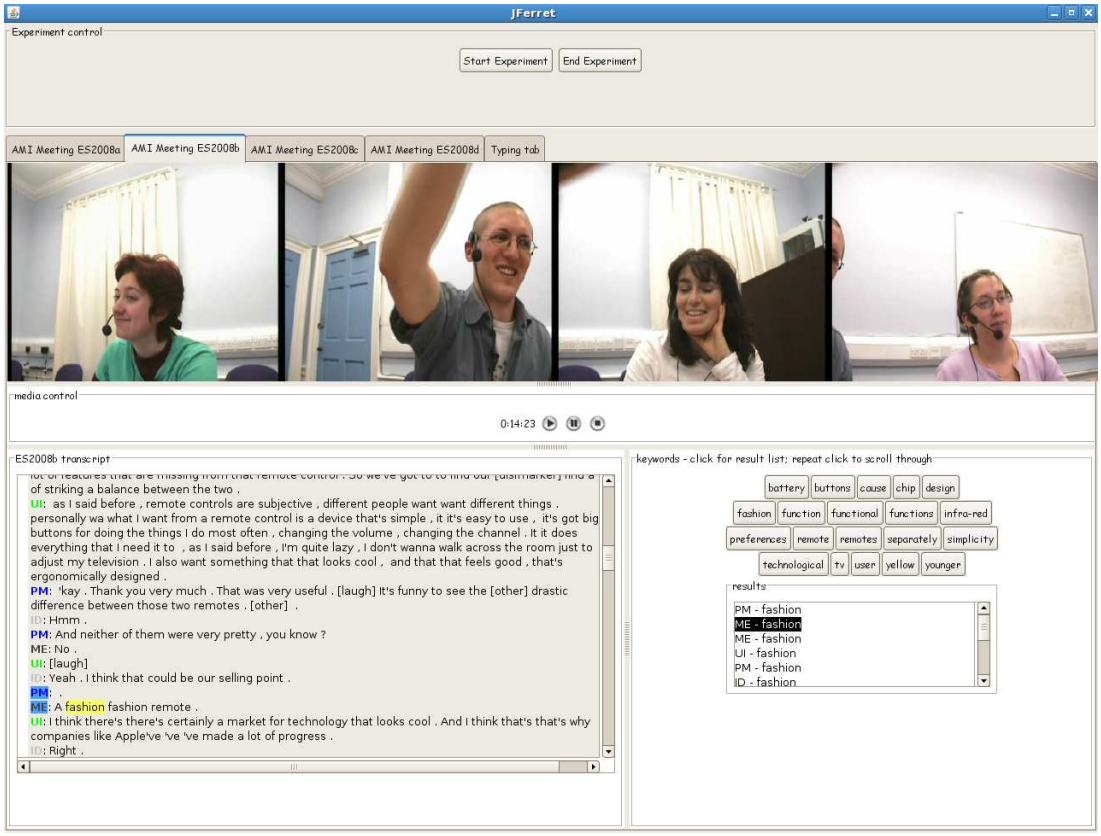


Figure 6.1: Condition KAM Browser

6.3.3 Browser Setup

The meeting browsers are built so as to exhibit as similar browser behaviour as possible across the experimental conditions. In other words, the interface is kept essentially the same in all conditions to eliminate any potential confounding factors relating to the user interface.

In each browser, there are 5 tabs for the 4 meetings and a writing pad. The writing pad is provided for the participant to author their decision audit summary. In each meeting tab, the videos displaying the 4 meeting participants are laid out horizontally with the media controls beneath. The transcript is shown in the lower left of the browser tab in a scroll window.

In Condition KAM, each meeting tab contains buttons corresponding to the top 20 keywords for that meeting. Pressing the button for a given keyword highlights the first instance of the keyword in the transcript, as well as opening a listbox illustrating all of the occurrences of the word in the transcript, giving the user a context in terms of the word's frequency. Subsequent clicks highlight the subsequent occurrences of the

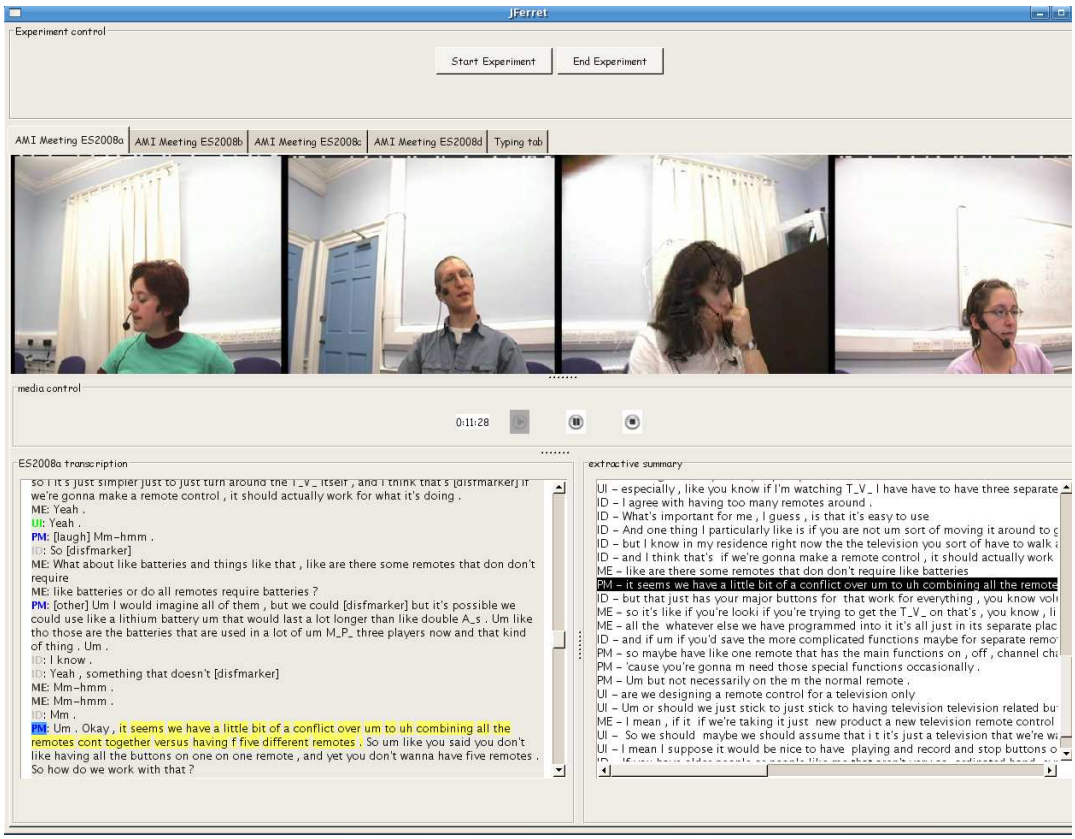


Figure 6.2: Conditions EAM and EAA Browser

word in the transcript, or the user may choose to navigate to keyword instances via the listbox.

In Conditions EAM and EAA, a scroll window containing the extractive summary appears next to the full meeting transcript. Clicking on any dialogue act in the extractive summary takes the user to that point of the meeting transcript and audio/video record.

In Conditions AMM and ASM, the abstractive summary is presented next to the meeting transcript. In Condition ASM, the abstractive summary has different tabs for *decision*, *problems*, *goals* and *actions*. Clicking on any abstract sentence highlights the first linked dialogue act in the transcript and also presents a listbox representing all of the transcript dialogue acts linked to that abstract sentence. The user can thus navigate either by repeatedly clicking the sentence, which in turn will take them to each of the linked dialogue acts in the transcript, or else they can choose a dialogue act from the listbox. The navigation options are underlyingly the same as Condition KAM. The primary difference between Conditions KAM, AMM and ASM on the one hand and Conditions EAM and EAA on the other is that the extractive dialogue acts link

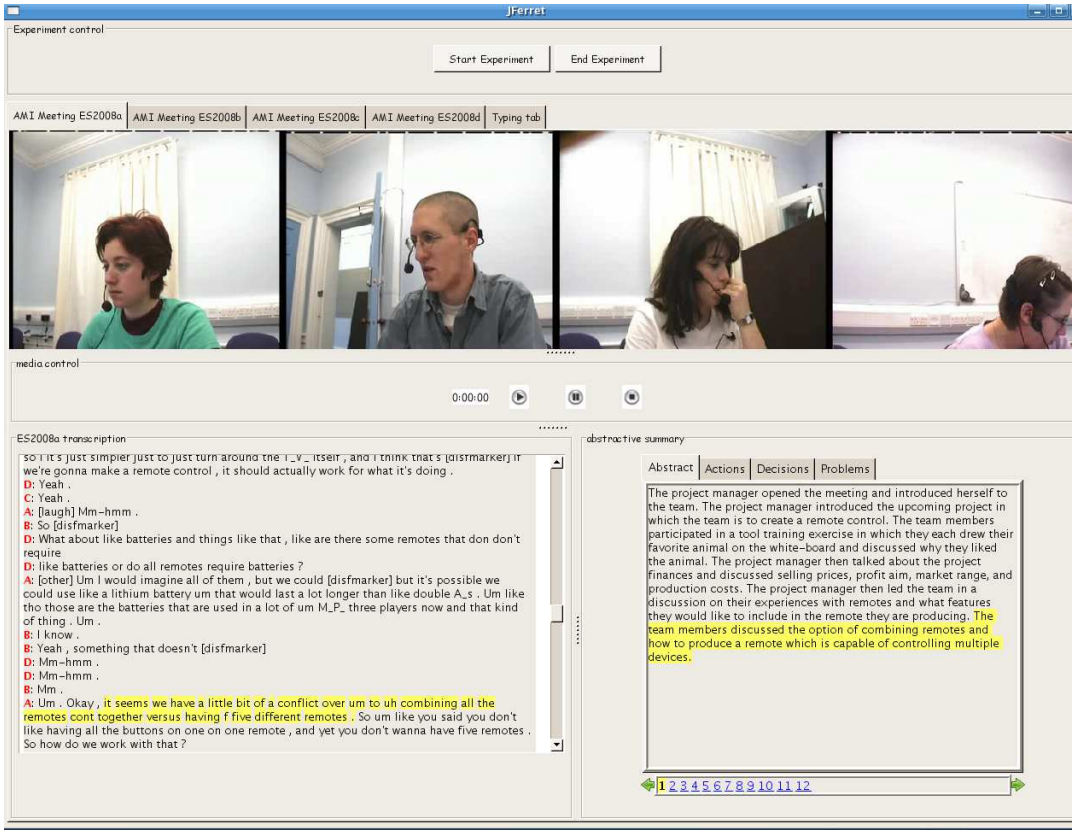


Figure 6.3: Condition AMM Browser

to only one point in the meeting transcript, whereas keywords and abstract sentences have multiple indices.

The browsers are designed in such a way that the writing tab where the participant types their answer is a fifth tab in addition to the four individual meeting tabs. As a consequence, the participant cannot view the meeting tabs while typing the answer; they are restricted to tabbing back and forth as needed. This was designed deliberately so as to be able to discern when the participant was working on formulating or writing the answer on the one hand and when they were browsing the meeting records on the other.

After reading the task instructions, each participant is briefly shown how to use the browser's various functions for navigating and writing in the given experimental condition. They are then given several minutes to familiarize themselves with the browser, until they state that they were comfortable and ready to proceed. The meeting used for this familiarization session is not one of the ES2008 meetings used in the actual task. In fact, it was one of the AMI non-scenario meetings; this is done so that the participant will not become familiar with the ES2008 meetings specifically or the

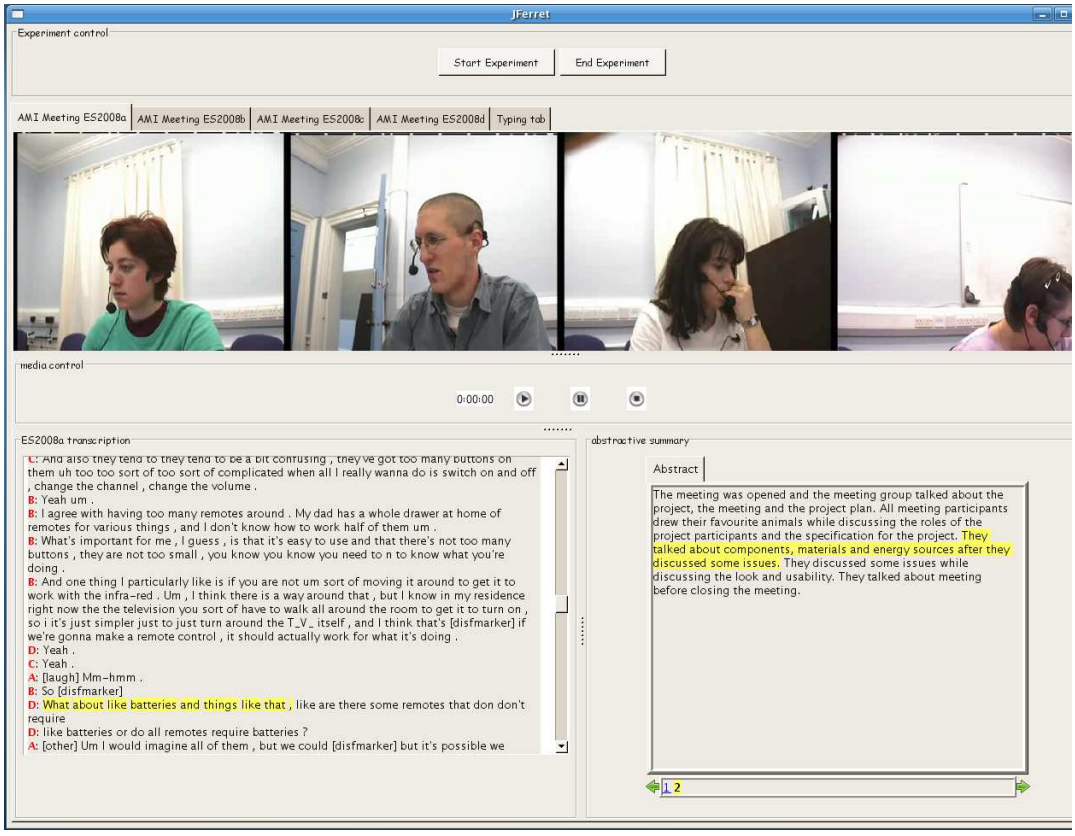


Figure 6.4: Condition ASM Browser

scenario meetings in general before beginning the task. This familiarization time is carried out before the task began so that we could control for the possibility that one condition would have a more difficult learning curve than the others.

6.3.4 Logfiles

In each condition of the experiment, we log a variety of information relating to the participant's browser use and typing. In all conditions, we log transcript clicks, media control clicks (i.e. play, pause, stop), movement between tabs, and characters entered into the typing tab, all of which are time-stamped. In Condition KAM, we log each keyword click and note its index in the listbox, e.g. the first occurrence of the word in the listbox. In Conditions EAM and EAA, each click of an extractive summary sentence is logged, and in the abstract conditions each abstract sentence click is logged along with its index in the listbox, analogous to the keyword condition. Because there are not multiple links in the extractive condition – in other words, each extract sentence links only to one transcript sentence – there is no need for listboxes and listbox indices.

To give an example, the following portion of a logfile from a Condition AMM task shows that the participant click on the transcript, played the audio, paused the audio, clicked link number 1 of sentence 5 in the Decisions tab for the given meeting, then switched to the typing tab and began typing the word "six."

```
2007-05-24T14:46:45.713Z transcript_jump 687.85 ES2008d.sync.1375
2007-05-24T14:46:45.715Z button_press play state media_d
2007-05-24T14:46:45.715Z button_press play state media_d
2007-05-24T14:47:30.726Z button_press pause state media_d
2007-05-24T14:47:30.726Z button_press pause state media_d
2007-05-24T14:47:52.379Z MASCOT (observation ES2008d): selected link
                                #1 in sentence #5 of tab 'decisions'
2007-05-24T14:47:53.613Z tab_selection Typing tab
2007-05-24T14:47:54.786Z typed_insert s 316
2007-05-24T14:47:54.914Z typed_insert i 317
2007-05-24T14:47:55.034Z typed_insert x 318
```

6.3.5 Evaluation Features

For evaluation of the decision audit task, there are three types of features to be analyzed: the answers to the users' post-questionnaires, human ratings of the users' written answers, and features extracted from the logfiles that relate to browsing and typing behaviour in the different conditions.

Upon completion of the decision audit task, we present each participant with a post-task questionnaire consisting of 10 statements with which the participant can state their level of agreement or disagreement via a 5-point Likert scale, such as *I was able to efficiently find the relevant information*, and two open-ended questions about the specific type of information available in the given condition and what further information they would have liked. Of the 10 statements evaluated, some are re-wordings of others with the polarity reversed in order to gauge the users' consistency in answering. See Appendix A (page 176) for the post-questionnaires in their entirety.

In order to gauge the goodness of a participant's answer, we enlist two human judges to do both *subjective* and *objective* evaluations. For the subjective portion, the judges first read through all 50 answers to get a view of the variety of answers. They then rate each answer using a 1-8 Likert-scale on criteria relating to the precision, recall and f-score of the answer. For the objective evaluation, three judges construct a gold-

Post-Questionnaire	Human Ratings	Logfile
Q1: *I found the meeting browser intuitive and easy to use*	*overall quality*	*task duration*
Q2: *I was able to find all of the information I needed*	*conciseness*	*first typing*
Q3: *I was able to efficiently find the relevant information*	*completeness*	*amount of tabbing*
Q4: *I feel that I completed the task in its entirety*	*task comprehension*	*perc. buttons clicked*
Q5: *I understood the overall content of the meeting discussion*	*participant effort*	*clicks per minute*
Q6: *The task required a great deal of effort*	*writing style*	*media clicks*
Q7: *I had to work under pressure*	*objective rating*	*click/writing correlation*
Q8: *I had the tools necessary to complete the task efficiently*	-	*unedited length*
Q9: *I would have liked additional information about the meetings*	-	*edited length*
Q10: *It was difficult to understand the content of the content of the meetings using this browser*	-	*num. meetings viewed*
Q11: -	-	*ave. writing timestamp*

Table 6.2: Decision Audit Evaluation Features

standard list of items that should be contained in an ideal summary of the decision audit (see Appendix C, page 182). For each participant answer, they check off how many of the gold-standard items are contained. Due to the fact that some participant answers included written text in paragraph form in addition to rough notes, summaries with both notes and text are evaluated twice, first considering all the text that was submitted and a second time considering only the written paragraphs were submitted. This is done because it was not clear whether the notes were meant to be submitted as part of the answer or were simply not deleted before time had expired. For this analysis we use only the full answers provided, however, to avoid putting ourselves in the position of trying to determine whether a participant did or did not intend to submit certain pieces of information.

The remainder of the features for evaluation are automatically derived from the logfiles. These features have to do with browsing and writing behaviour as well as the duration of the task. These include the total experiment length, the amount of time before the participant began typing their answer, the total amount of tabbing the user did normalized by experiment length, the number of clicks on content buttons (e.g. keyword buttons or extractive summary sentences) per minute, the number of content button clicks normalized by the number of unique content buttons, number of times the user played the audio/video stream, the number of content clicks prior to the user clicking on the writing tab to begin writing, the document length including deleted characters, the document length excluding deleted characters, how many of the four

meetings the participant looked at, and the average typing timestamp normalized by the experiment length.

The total experiment length is included because it is assumed that participants would finish earlier if they had better and more efficient access to the relevant information. The amount of time before typing begins is included because it is hypothesized that efficient access to the relevant information would mean that the user would begin typing the answer sooner. The total amount of tabbing is considered because a participant who is tabbing very often during the experiment is likely jumping back and forth between meetings trying to find the information, indicating that the information is not conveniently indexed. The content clicks are considered because a high number of clicks per minute would indicate that the participant is finding that method of browsing to be helpful, and the number of content clicks normalized by the total unique content buttons indicates whether they made full use of that information source. The number of audio/video clicks is interesting because it is hypothesized that a user without efficient access to the relevant information will rely more heavily on scanning through the audio/video stream in search of the answers. The number of content clicks prior to the user moving to the writing tab indicates whether a content click is helpful in finding a piece of information that led to writing part of the answer. The document length is considered because a user with better and more efficient access to the meeting record will be able to spend more time writing and less time searching. Because the logfiles show deleted characters, we calculate both the total amount of typing and the length of the final edited answer in characters. The number of meetings examined is considered because a user who has trouble finding the relevant information may not have time to look at all four meetings. The final feature, which is the average timestamp normalized by the experiment length, is included because a user with efficient access to the information will be able to write the answer throughout the course of the experiment, whereas somebody who has difficulty finding the relevant information may try to write everything at the last available moment.

Table 6.2 lists all of the features used for evaluation.

6.4 Results

The following sections present the post-questionnaire results, the human subjective and objective evaluation results, and the analysis of browsing behaviours.

Question	KAM	EAM	EAA	AMM	ASM
Q1: *I found the meeting browser intuitive and easy to use*	3.8	4.0	3.0^{2}_{AMM}	$4.3^{EAA,ASM}$	3.7_{AMM}
Q2: *I was able to find all of the information I needed*	2.9_{AMM}	3.8	2.9_{AMM}	$4.1^{KAM,EAA,ASM}$	3.0_{AMM}
Q3: *I was able to efficiently find the relevant information*	2.8_{AMM}	3.4^{ASM}	2.5_{AMM}	$4.0^{KAM,EAA,ASM}$	$2.65_{EAM,AMM}$
Q4: *I feel that I completed the task in its entirety*	2.3_{AMM}	3.1	2.3	3.2^{KAM}	2.9
Q5: *I understood the overall content of the meeting discussion*	3.8	**4.5**	3.9	4.1	3.9
Q6: *The task required a great deal of effort*	3.0	2.6^{EAA}	3.9_{EAM}	3.1	3.2
Q7: *I had to work under pressure*	3.3	**2.6**	3.3	2.7	3.1
Q8: *I had the tools necessary to complete the task efficiently*	3.1_{EAM}	$4.3^{KAM,EAA,ASM}$	3.0_{EAM}	4.1	3.5_{EAM}
Q9: *I would have liked additional information about the meetings*	3.0_{EAM}	2.0^{KAM}	2.4	2.6	2.7
Q10: *It was difficult to understand the content of the meetings using this browser*	2.1	$1.5^{EAA,ASM}$	2.7_{EAM}	2.0	2.3_{EAM}

Table 6.3: Post-Questionnaire Results

For each score in the table, that score is significantly better than the score for any conditions in superscript, and significantly worse than the score for any condition in subscript.

6.4.1 Post-Questionnaire Results

Table 6.3 gives the post-questionnaire results for each condition. For each score in the table, that score is significantly better than the score for any conditions in superscript, and significantly worse than the score for any condition in subscript. The only significant results listed are those that are significant at the level ($p<0.05$) according to non-paired t-test. Results that are not significant but are nonetheless unexpected or interesting are listed in boldface.

Question 1 For the first post-questionnaire question, *I found the meeting browser intuitive and easy to use*, the best condition overall is Condition AMM, incorporating human abstracts, followed by Condition EAM. There is no significant difference between the two conditions. The lowest score is for Condition EAA. Since the only difference between Conditions EAM and EAA is manual versus ASR transcripts, it's clear that ASR alone makes the browser less straight-forward and easy to use for participants.

Question 2 For the second post-questionnaire question, *I was able to find all of the information I needed*, the conditions roughly form two groups. Conditions AMM

and EAM are again at the top, scoring 4.1 and 3.8 respectively, while the remaining three conditions all score around 3.0. There is no significant difference between Conditions AMM and EAM.

Question 3 The third question was *I was able to efficiently find the relevant information*, and for this criterion the human abstracts are clearly superior, performing significantly better than Conditions KAM, EAA and ASM. Condition EAM is second best and not significantly worse than Condition AMM, but is substantially lower on average. Surprisingly, the *automatic* abstracts perform worse than the baseline Condition KAM on this criterion.

Question 4 For question four, *I feel that I completed the task in its entirety*, the scores overall are somewhat low, indicating the difficulty of the task. The best conditions are Condition EAM and Condition AMM with scores of 3.1 and 3.2 respectively. Condition AMM is significantly better than the baseline Condition KAM. The lack of large differences across conditions regarding this criterion confirms that it is a challenging task to complete in the allotted time.

Question 5 For question five, *I understood the overall content of the meeting discussion*, the best condition is Condition EAM, extractive on manual transcripts, with a score of 4.5. While this is several points higher than even the human abstract condition, there are no significant differences between the conditions for this criterion. Nonetheless, it is very encouraging that the extractive conditions provide a good overview of the meeting content compared with the other conditions. Even with ASR, Condition EAA fares very well on this criterion.

Question 6 For question six, *The task required a great deal of effort*, Condition EAM is again the best with a score of 2.6 (the lower the score, the better). The worst score, i.e. the highest, is Condition EAA, showing that an ASR transcript does increase the effort required to complete the task compared with having a manual transcript.

Question 7 Similarly for question seven, *I had to work under pressure*, Condition EAM is the best with a score of 2.6 and Condition AMM is comparable with a score of 2.7. There are no significant differences between the conditions. Conditions KAM and EAA score the worst on this criterion. This result shows that extractive summaries can allow users to make efficient use of their time, and

that the presence of errorful ASR transcripts increases the participants' sense of being under pressure to complete the task.

Question 8 For question eight, *I had the tools necessary to complete the task efficiently*, Condition EAM is again the highest with a score of 4.3 followed by Condition AMM with a score of 4.1. Condition EAM is significantly better on this criterion than Conditions KAM, EAA and ASM. This is quite an encouraging result for extractive summarization, as the question directly addresses the tools available to the user and the extractive condition comes out on top. Not only does it perform the best overall, but the score of 4.3 is quite high on the 1-5 Likert scale, indicating user satisfaction with the browser content. From this criterion, we also find that the presence of errorful ASR transcripts decreases user satisfaction with the tools provided.

Question 9 For the final two questions, Condition EAM again performs the best. For the question *I would have liked additional information about the meetings*, Condition EAM is rated with a 2.0 on average, followed by Condition EAA with a score of 2.4 Thus, the two extractive conditions come out on top, superior to even the human abstract condition.

Question 10 For the question *It was difficult to understand the content of the meetings using this browser*, Condition EAM is rated with a 1.5 on average followed by Condition AMM with an average score of 2.0 (again, the lower the better for the last two questions). For this criterion, Condition EAM is considerably better than the rest, with significant results compared with Conditions EAA and ASM. The low score for Condition EAA shows that the incorporation of ASR transcripts does make it more difficult to understand the meetings for participants in this task, but even that score of 2.7 for Condition EAA is not as high on the Likert scale as might be expected. These final two questions indicate that users are quite satisfied with the information provided by the extractive summaries and that the summaries allow them to understand the meetings without much difficulty.

6.4.1.1 Discussion

It can first be noted that participants in general find the task to be challenging, as evidenced by the average answers on questions 4, 6 and 7. The task was designed to be challenging and time-constrained, because a simple task with a plentiful amount of

allotted time would allow the participants to simply read through the entire transcript or listen and watch the entire audio/video record in order to retrieve the correct information, disregarding other information sources. The task as designed requires efficient navigation of the information in the meetings in order to finish the task completely and on time.

The gold-standard human abstracts were rated highly on average by participants in that condition. Judging from the open-ended questions in the post-questionnaire, people found the summaries and specifically the summary subsections to be very valuable sources of information. One participant remarked "Very well prepared summaries. They were adequate to learn the jist [sic] of the meetings by quickly skimming through... I especially liked the tabs (Decisions, Actions, etc.) that categorised information according to what I was looking for." As mentioned earlier, this gold-standard condition was expected to do particularly well considering that it is a decision audit task and the abstractive summaries contain subsections that are specifically focused on decision-making in the meetings.

The results of the post-questionnaire data are quite encouraging in that the users seem very satisfied with the extractive summaries relative to the other conditions. It is not surprising that the gold-standard human-authored summaries are ranked best overall on several criteria, but even on those criteria the extractive condition on manual transcripts is a close second. For question 5, which relates to overall comprehension of the information in the meetings, extractive summaries are rated the highest of all. Extractive summaries of manual transcripts are also rated the best in terms of the effort required to conduct the task. But perhaps the most compelling result is on question 8, relating to having the tools necessary to complete the task. Not only is Condition EAM rated the best, but it is *significantly better* than all conditions except the gold-standard human abstracts. These results taken together indicate that extractive summaries are natural to use as navigation tools, that they facilitate understanding of the meeting content, and allow users to be more efficient with their time. From the viewpoint of user satisfaction, this result is the best that could be hoped for.

However, it is quite clear that the errors within an ASR transcript present a considerable problem for users trying to quickly retrieve information from the meetings. While it has repeatedly been shown that ASR errors do not cause problems for our algorithms according to intrinsic measures (Chapters 4 and 5), these errors make user comprehension more difficult. For the questions relating to the effort required, the tools available, and the difficulty in understanding the meetings, Condition EAA is

easily the worst, scoring even lower than the baseline condition. It should be noted however, that a baseline such as Condition KAM is not a true baseline in that it is working off of *manual* transcripts and would be expected to be worse when applied to ASR. As mentioned earlier, the baseline is a challenging baseline in that respect. Judging from the open-ended questions in the post-questionnaires, it's clear that at least two participants found the ASR so difficult to work with that they tended not to use the extractive summaries, let alone the full transcript, relying instead on watching the audio/video as much as possible. For example, one person responded to the question "How useful did you find the list of important sentences from each meeting?" with the comment "Not at all, because the voice recognition technology did not work properly. The only way to understand the discussion was to listen to it all sequentially, and there simply wasn't time to do that." We will analyze users' browsing behaviour in much more detail below. Here we give a brief excerpt from the ASR summary for the fourth meeting, illustrating the difficulty posed by errorful dialogue acts:

> **Speaker D:** Could the middle button on the on screen menu function of the power button?
>
> **Speaker B:** And then finally we have Um the martian or the pair yeah right.
>
> **Speaker B:** Oh it's a bit different a little bit more of a creative feel.
>
> **Speaker B:** Are you have the on off Foucault stammer on the top.
>
> **Speaker B:** You have your channel changing volume changing buttons and your menu button right here in the middle.

These findings regarding the difficulty of human processing of ASR transcripts will change and improve as the state-of-the-art in speech recognition improves. The finding also indicates that the use of confidence scores in summarization is desirable. While summarization systems naturally tend to extract units with lower WER, the summaries can likely be further improved for human consumption by compression via the filtering of low-confidence words.

6.4.2 Human Evaluation Results - Subjective and Objective

6.4.2.1 Subjective Evaluation

Table 6.4 gives the results for the human subjective and objective evaluations. For each score in the table, that score is significantly better than the score for any conditions in superscript, and significantly worse than the score for any condition in subscript.

The only significant results listed are those that are significant at the level ($p<0.05$). Results that are not significant but are nonetheless unexpected or interesting are listed in boldface.

Before beginning the subjective evaluation of decision audit answers, the two human judges read through all 50 answers in order to gauge the variety of answers in terms of completeness and correctness. They then rate each answer on several criteria roughly related to ideas of precision, recall and f-score, as well as effort, comprehension and writing style. They use a 1-8 Likert scale for each criterion. We then average their scores to derive a combined score for each criterion. Both judges are researchers at DFKI and neither is the author of this thesis, but both are very familiar with the AMI corpus data.

Question 1 For the "overall quality" criterion, Condition AMM, incorporating human abstracts, is superior, with an average of 4.85. The worst conditions overall are Condition KAM and Condition EAA, each scoring around 3.0. Extracts of manual transcripts and automatic abstracts are slightly worse than the gold-standard condition. Condition ASM is rated second best, after AMM.

Question 2 For the evaluation of "conciseness," the trends are largely the same as for the "overall quality" question. Condition AMM is the best with an average of 4.85, followed by Conditions 4 and 1 with scores of 4.45 and 4.25, respectively. Condition KAM is easily the worst, performing significantly worse than every other condition with the exception of Condition EAA.

Question 3 The pattern is similar for the evaluation of "completeness," with Condition AMM faring best of all followed by Conditions ASM and EAM in order. On this criterion there is a clearer gap between the gold-standard condition and the remaining conditions, illustrating the utility of a manual abstract for providing complete coverage of the meeting. Worst for "completeness" is Condition KAM.

Questions 4 and 5 For the criteria of "task comprehension" and "participant effort", we find Condition EAM scoring nearly as well as Condition AMM. For Condition EAA, incorporating ASR, these scores significantly decrease, illustrating the challenge that an errorful transcript poses in terms of users understanding the task and demonstrating a concerted effort to satisfy the information need. Of course, it is difficult to discern incomprehension or low effort from what could simply be a difficult task.

Criterion	KAM	EAM	EAA	AMM	ASM
Q1: *overall quality*	3.0_{AMM}	4.15	3.05_{AMM}	$4.65^{KAM,EAA}$	4.3
Q2: *conciseness*	$2.85_{EAM,AMM,ASM}$	4.25^{KAM}	3.05_{AMM}	$4.85^{KAM,EAA}$	4.45^{KAM}
Q3: *completeness*	2.55_{AMM}	3.6	2.6_{AMM}	$4.45^{KAM,EAA}$	3.9
Q4: *task comprehension*	$3.25_{EAM,AMM}$	$\mathbf{5.2}^{KAM,EAA}$	$3.65_{EAM,AMM}$	$5.25^{KAM,EAA}$	4.7
Q5: *participant effort*	4.4	$\mathbf{5.2}^{EAA}$	$3.7_{EAM,AMM,ASM}$	5.3^{EAA}	4.9^{EAA}
Q6: *writing style*	4.75	5.65^{EAA}	$4.1_{EAM,AMM,ASM}$	5.7^{EAA}	5.8^{EAA}
Q7: *objective rating*	4.25_{AMM}	7.2	5.05_{AMM}	$9.45^{KAM,EAA}$	7.4

Table 6.4: Human Evaluation Results - Subjective and Objective

For each score in the table, that score is significantly better than the score for any conditions in superscript, and significantly worse than the score for any condition in subscript.

Question 6 For the evaluation of "writing style", we find that Conditions EAM, AMM and ASM are rated similarly, while Condition EAA scores the worst. There may be numerous factors for how ASR affects writing style in this task, but it may be that users are unable to decipher exactly what is discussed and subsequently their write-ups reflect this partial understanding, or it could simply be that they have less time to spend on writing because their browsing is less efficient. We will examine this latter point in further detail in the logfile results section below.

What these findings together help illustrate is that extractive summaries can be very effective for conducting a decision audit by helping the user to generate a concise, complete high-quality answer, but that the introduction of ASR has a measurable and significant impact on the subjective evaluation of quality. Interestingly, the scores on each criterion and for each condition tend to be somewhat low on the Likert scale, due to the difficulty of the task.

6.4.2.2 Objective Evaluation

After the annotators carried out their objective evaluations, they met again and went over all experiments where their ratings diverged by more than two points, in order to form a truly *objective* and agreed-upon evaluation of how many gold-standard items each participant found. There were 12 out of 50 ratings pairs that needed revision in this manner. After the judges' consultation on those 12 pairs of ratings, each experiment was given a single objective rating. The judges mentioned that they found this portion of the evaluation much more difficult than the subjective evaluations, as there was often ambiguity as to whether a given answer contained a given gold-standard item or not.

According to the objective evaluation, Condition AMM is superior, with an average more than two points higher than the next best condition. The worst overall is the baseline Condition KAM, averaging only 4.25 hits. However, while the worst two conditions are significantly worse than the best overall condition, there are no significant differences between the other pairs of conditions, e.g. Condition EAA incorporating ASR is not significantly worse than Conditions EAM and ASM. So even with an errorful transcript, participants in Condition EAA are able to retrieve the relevant pieces of information at a rate not significantly worse than participants with a manual transcript. The quality may be worse from a subjective standpoint, as evidenced in the previous section, but the decision audit answers are still informative and relevant.

For the objective evaluation, in any given condition there is a large amount of variance that is simply down to differences between users. For example, even in the gold-standard Condition AMM there are some people who can only find one or two relevant items whilst others find 16 or 17. Given a challenging task and a limited amount of time, some people may have simply felt overwhelmed in trying to locate the informative portions efficiently.

Table 6.4 summarizes the human evaluation results for both the subjective and objective criteria.

6.4.2.3 Discussion

For the objective human evaluation, the gold-standard condition scores substantially higher than the other conditions in hitting the important points of the decision process being audited. This goes to show that there is much room for improvement in terms of automatic summarization techniques. However, Conditions EAM, EAA and ASM average much higher than the baseline Condition KAM. There is considerable utility in such automatically-generated documents. It can also be noted that Condition EAM is the best of the conditions with fully-automatic content selection (Condition ASM is not fully automatic).

Perhaps the most interesting result of the objective evaluation is that Condition EAA, which uses ASR transcripts, does not deteriorate relative to Condition EAM as much as might have been expected considering the post-questionnaire results. What this seems to demonstrate is that ASR errors are annoying for the user but that the users are able to look past the errors and still find the relevant information efficiently. Condition EAA scores much higher than the baseline Condition KAM that utilizes *manual* transcripts, and this is a powerful indicator that summaries of errorful documents are

still very valuable documents.

An interesting question is whether participants' self-ratings on task performance correlate with their actual objective performance according to the human judges. To answer this question, we calculate the correlation between the scores from post-questionnaire Q4 and the objective scores. The statement Q4 from the post-questionnaire is "I feel that I completed the task in its entirety." The result is that there is a moderate but significant positive correlation between participant self-ratings and objective scores (pearson=0.39, $p<0.005$).

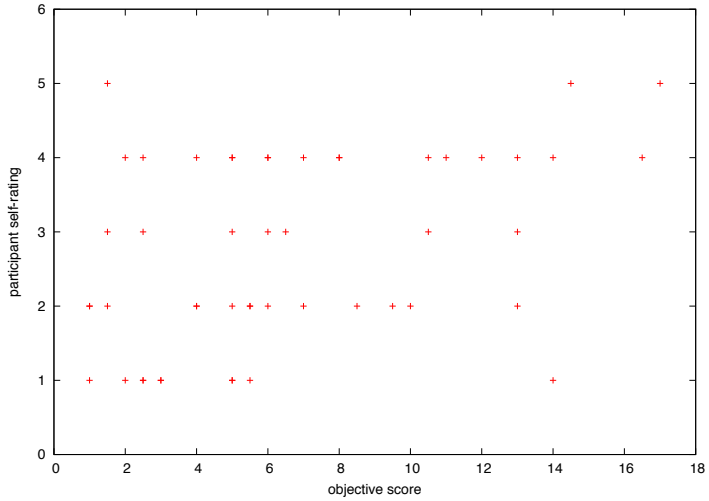


Figure 6.5: Objective Scores and Post-Questionnaire Scores

Figure 6.5 shows the relationship between the objective ratings and participant self-ratings for all 50 participants. While the positive correlation is evident, an interesting trend is that while there are relatively few people who score highly on the objective evaluation but score low on the self-ratings, there are a fair number of participants who have a low objective score but rate themselves highly on the post-questionnaire. A challenge with this type of task is that the participant simply may not have a realistic idea of how much relevant information is out there. After retrieving four or five relevant items, they may feel that they've completed the task entirely. This result is similar to the finding by Whittaker et al. (2008), mentioned in the discussion of previous work, where participants often feel that they performed better than they really did.

6.4.3 Extrinsic/Intrinsic Correlation

In order to determine whether available intrinsic evaluation metrics predict the discrepancy in ratings between manual and ASR transcripts, we score the extractive sum-

Metric	Man	ASR
Objective	7.2	5.05
PQ4	3.1	2.4
ROUGE-2	0.55	0.41
ROUGE-SU4	0.57	0.47
Weighted F	0.48	0.46

Table 6.5: Comparison of Extrinsic/Intrinsic Scores

Feature	KAM	EAM	EAA	AMM	ASM
Q1: *duration*	45.4	43.1	45.4	45.42	43.2
Q2: *first typing*	16.25	13.9	17.14	8.61	10.22
Q3: *tabbing*	0.98	0.81^{AMM}	0.72^{AMM}	$1.4_{EAM,EAA}$	1.13
Q4: *perc. buttons clicked*	0.39	0.11	0.08	0.08	0.18
Q5: *clicks per minute*	1.33	**2.24**	1.47	1.99	0.83
Q6: *media clicks*	15.4^{EAA}	14.4^{EAA}	$40.4_{KAM,EAM,AMM}$	16.6^{EAA}	20.6
Q7: *click/writing corr.*	0.03	0.01	0.01	0.01	0.01
Q8: *unedited length*	1400	1602	1397	**2043**	1650
Q9: *edited length*	1251	1384	1161	**1760**	1430
Q10: *num. meetings*	3.9	4.0	3.9	4.0	4.0
Q11: *ave. writing timestamp*	0.68	0.73	$0.76_{AMM,ASM}$	0.65^{EAA}	0.65^{EAA}

Table 6.6: Logfile Feature Results

For each score in the table, that score is significantly better than the score for any conditions in superscript, and significantly worse than the score for any condition in subscript.

maries in both conditions using ROUGE and weighted f-score. For the ROUGE evaluation, gold-standard human extracts are used as the reference summaries (multiple human abstracts are lacking for this particular meeting set). ROUGE is run with the standard DUC parameters. Figure 6.5 shows the results of these intrinsic evaluations along with the objective human results and post-questionnaire statement Q4, "I feel that I completed the task in its entirety." All metrics do show a decline on ASR compared with manual transcripts for these four meetings. The difference in scores is most pronounced with ROUGE-2, while weighted f-score shows the least decline on ASR. This is likely due to the fact that ROUGE evaluations are carried out at the n-gram level while weighted f-score works only at the dialogue act level. Weighted f-score does not directly take ASR errors into account; the impact of ASR is on whether or not the error-filled dialogue acts are extracted in the first place.

6.4.4 Logfile Results

Table 6.6 gives the results for the logfiles evaluation. For each score in the table, that score is significantly better than the score for any conditions in superscript, and significantly worse than the score for any condition in subscript. The only significant results listed are those that are significant at the level ($p<0.05$). Results that are not significant but are nonetheless unexpected or interesting are listed in boldface.

Feature 1 One result that was not anticipated is that almost all participants take the full 45 minutes to complete the experiment. There are no significant differences between the conditions on this criterion, though Condition EAM has the lowest average task duration at 43 minutes. One hypothesis is that paid volunteers want to do as thorough of a job as possible and so remain for the entirety of the allotted time even if they have finished the bulk of the experiment earlier. This is backed anecdotally by participants reporting afterwards that "you can always use more time," suggesting that answers can always be refined even when near completion. More generally, it turned out to be a challenging task to complete in 45 minutes, regardless of condition. In hindsight, it perhaps would have been better to provide a longer amount of time in the hope that differences between conditions would become more evident in terms of task duration.

Feature 2 The second feature is the amount of time before the participant began typing the answer. Condition AMM is best overall with an average time of 8.6 minutes. Condition ASM is next best with 10.225 minutes, Condition EAM with 13.9 minutes, Condition KAM with 16.25 minutes and Condition EAA with 17.137 minutes. However, there are no significant differences between conditions. It is nonetheless clear that human abstracts allow the users to quickly index into the relevant portions of the meeting and begin writing the decision audit answer quite quickly.

Feature 3 The results of the third feature are surprising. The metric is the total amount of moving between browser tabs, normalized by the length of the experiment. The intuition behind the inclusion of this feature is that users who have efficient access to the relevant, important information will not need to continually tab back and forth between the browser tabs, searching for the information. The best (i.e. lowest) score overall is Condition EAA, extractive summaries on ASR transcripts, followed by Condition EAM, extractive summaries on manual tran-

scripts. The worst overall is Condition AMM, human abstracts. Conditions EAM and EAA are significantly better than Condition AMM.

Features 4 and 5 The fourth and fifth features relate to the number of clicks on content items, e.g. keyword clicks or extractive summary clicks. The fourth feature normalizes the number of clicks by the total number of content buttons. For example, if five unique keyword buttons were clicked out of a possible 20, the score would be 0.25. The fifth feature normalizes the number of content clicks by the length of the experiment, i.e. it represents the number of clicks per minute. For the fourth feature, Condition KAM is the best overall with an average score of 0.386, significantly better than Conditions EAA and AMM. For the fifth feature, Condition EAM is best overall with an average of 2.24 content clicks per minute, followed by Condition AMM with an average of 1.993. Condition ASM is the worst with an average of 0.831. There are no significant differences between conditions. The fifth logfile feature is more likely to be reliable than the fourth, as the number of keywords for each meeting is only 20 and it's not surprising that the percentage of buttons clicked is higher than for the other conditions. The clicks-per-minute result is interesting for two reasons: extracts are used for navigation with considerably more frequency than the other conditions, and there are very few navigation clicks in Condition ASM, incorporating automatic abstracts. We find that with extracts on ASR, users click the extracted dialogue acts less often than on manual transcripts, but still more often than in Conditions KAM and ASM.

Feature 6 The sixth feature is the number of media clicks, i.e. the number of times the user played the audio/video. The best condition is Condition EAM, followed by Condition KAM. The most interesting and dramatic result, however, is that Condition EAA, extractive summarization on ASR, is much worse than all the other conditions. Whereas the average number of media clicks for Condition EAM is 14.4, for Condition EAA it is 40.4. This illustrates that the errorful ASR transcripts cause the users to rely much more heavily on the audio/video stream. Participants in Condition ASM also rely more on the audio/video streams than participants in the top three conditions.

Feature 7 The seventh feature is the proximity of content clicks to writing tab clicks. Condition KAM is best overall, but there are no significant differences between conditions. It seems to simply be a rare occurrence for a user to click a content

item and began writing soon afterwards. More likely, they click a content item and navigate to that part of the meeting, study the transcript in more detail, and finally synthesize the information in the writing tab.

Features 8 and 9 The eighth and ninth features relate to the length of the user's answer. For feature eight, the unedited answer length, Condition AMM is best overall with an average character length of 2043.2. The worst is the baseline Condition KAM with an average of 1399.6. Interestingly, for the ninth feature - edited answer length - the scores are much closer. Condition AMM is still the best overall with an average length of 1760.6, but Condition KAM is 1251.1. This illustrates that users in Condition AMM have much more time for editing and refining their answers. They might begin by writing everything they find that seems relevant, then they condense or combine information for the final answer.

Feature 10 The tenth feature is the number of meetings the user looked at. The intuition is that if a given condition is not very efficient in the way that it presented information, users might not have time to look at all the data. In reality, however, almost all participants looked at all of the meetings, and so there are no differences on this criterion.

Feature 11 The final feature is the average location within the 45 minute period of the user typing. That is, it is the average of the timestamps normalized by the initial timestamp. The intuition is that users in a condition with more efficient access to information will do more typing early on in the experiment, whereas a person in a condition with an inefficient browser would be forced to do much of the writing at the end of the experiment. Condition AMM was best overall with a score of 0.650, whereas Condition EAA was the worst with a score of 0.725. Participants with access to a human summary are able to do the bulk of their writing earlier on in the experiment, whereas participants using an ASR transcript do much of their writing towards the end of the experiment. In the latter case, this leaves them less time for revision, which is presumably related to the low writing quality scores presented in the previous section on subjective evaluations.

6.4.4.1 Discussion

It is difficult to derive a single over-arching conclusion from the logfile results, but there are several interesting results on specific logfile features. Perhaps the most interesting is the dramatic difference that exists in terms of relying on the audio/video record when using ASR. The average number of media clicks when using extractive summaries on manual transcripts is only just above 14, but when applied to ASR this number is over 40 clicks. This ties together several interesting results from the post-questionnaire data, the human evaluation data, and the logfile data. While the ASR errors seem to annoy the participants and therefore affect their user satisfaction ratings, they are nonetheless able to employ the ASR-based summaries to locate the relevant information efficiently and thereby score highly according to the human objective evaluation. Once they have indexed into the meeting record, they then rely heavily on the audio/video record presumably to disambiguate the dialogue act context. It is *not* the case that participants in this condition used only the audio/video record and disregarded the summaries, as they clicked the content items more often than in Conditions KAM and ASM (Q5). Overall, the finding is thus that ASR errors are annoying but do not obscure the value of the extractive summaries.

It is also interesting that both extractive conditions lead to participants needing to move between meeting tabs less than in other conditions. As mentioned above, the intuition behind the inclusion of this feature was that a lower number would be better because it meant the user was finding information efficiently. However, it's surprising that Condition EAA scored the "best" and Condition AMM the "worst." It may be the case that participants in Condition AMM felt more free to jump around because navigation was generally easier.

Many of the logfile features confirm that the human abstract gold-standard is difficult to challenge in terms of browsing efficiency. Users in this condition begin typing earlier, write most of their answer earlier in the task, write longer answers, and have more time for editing.

6.5 General Discussion

Overall these results are very good news for the extractive summarization paradigm. Users find extractive summaries to be intuitive, easy-to-use and efficient, are able to employ such documents to locate the relevant information in a timely manner accord-

ing to human evaluations, and users are able to adapt their browsing strategies to cope with ASR errors. While extractive summaries might be far from what people conceptualize as a meeting summary in terms of traditional meeting minutes, they are intuitive and useful documents in their own right.

Specifically, we have found that users in Condition EAM are very satisfied with the tools at their disposal, with the efficiency and intuitiveness of the browser setup, and their ability to rapidly find the relevant information. Condition EAM is the superior condition for several post-questionnaire criteria, such as Q8, which asks whether the user has the tools necessary to find the relevant information efficiently. In Condition EAA, incorporating ASR, users reported that they understood the overall content of the meeting discussions and did not desire any additional information, giving positive ratings compared with other conditions. The ASR did, however, affect their efficiency and ease-of-use ratings.

For the subjective human evaluation, the gold-standard Condition AMM was rated the best on nearly all criteria, but was challenged by Condition EAM on several of them, including the criteria of task comprehension and participant effort. Condition EAM also had high scores for overall quality, conciseness and completeness compared with Condition AMM. While the answers in Condition EAA were scored more severely in the subjective evaluation, the human *objective* evaluation showed that participants working with ASR were still able to locate the relevant pieces of information at a rate not significantly worse than participants using manual transcript extracts.

Finally, there are a couple of especially interesting results from the logfiles analysis. First of all, participants in Condition AMM are able to answer the question earlier in the experiment than participants in Condition EAA. Second, participants in Condition EAA rely much more on the audio/video streams than participants in other conditions.

Perhaps the most interesting result from the decision audit overall is regarding the effect of ASR on carrying out such a complex task. While participants using ASR find the browser to be less intuitive and efficient, they nonetheless feel that they understand the meeting discussions and do not desire additional information sources. In a subjective human evaluation, the quality of the answers in Condition EAA suffers according to most of the criteria, including writing style, but the participants are still able to find many of the relevant pieces of information according to the objective human evaluation. We find that users are able to adapt to errorful transcripts by using the summary dialogue acts as navigation and then relying much more on audio/video for disambiguating the conversation in the dialogue act context. Extractive summaries,

even with errorful ASR, are useful tools for such a complex task, particularly when incorporated into a multi-media browser framework.

There is also the possibility of creating browsing interfaces that minimize the user's direct exposure to the ASR transcript. Since we found in Chapters 4 (page 48) and 5 (page 81) that ASR does not pose a problem for our summarization algorithms, we could locate the most informative portions of the meeting and present the user with edited audio and video and limited or no textual accompaniment, to give one example.

Further regarding how one might minimize the impact of ASR errors, an interesting study would be to have human annotators perform extractive summarization on ASR transcripts, rather than our current method of mapping such annotations from manual transcripts on to ASR transcripts. It might be the case that humans would select markedly different subsets of dialogue acts from an ASR transcript than they would from a manual transcript, and studying these differences could inform future work on automatic extractive summarization of ASR output.

6.6 Conclusion

We have presented an extrinsic evaluation paradigm for the automatic summarization of spontaneous speech in the meetings domain: a decision audit task. This represents the largest extrinsic evaluation of speech summarization to date. In each condition of the experiment, users were able to utilize the derived content in order to find and extract information relevant to a specific task need. The largely positive results for the extractive conditions justify continued research on this summarization paradigm. However, the considerable superiority of gold-standard abstracts in many respects also support the view that research should begin to try to bridge the gap between extractive and abstractive summarization (Kleinbauer et al., 2007). In Chapter 7 (page 126) we present work relevant to that challenge.

It is widely accepted in the summarization community that there should be increased reliance on extrinsic measures of summary quality. It is hoped that the decision audit task will be a useful framework for future evaluation work. For development purposes, it is certainly the case that intrinsic measures are indispensable: as mentioned before, in this work we use intrinsic measures to evaluate several summarization systems against each other and use extrinsic measures to judge the usefulness of the extractive methods in general. Intrinsic and extrinsic methods should be used hand-in-hand, with the former as a valuable development tool and predictor of useful-

ness and the latter as a real-world evaluation of the state-of-the-art.

Chapter 7

The Extractive-Abstractive Continuum: Meta Comments in Meetings

7.1 Introduction

The vast majority of automatic summarization work on both speech and text to date has been *extractive* in nature. The reasons are that such techniques are domain-independent, do not require a deep understanding of the source document(s), and do not require a generation component. Jones (1999) has described the summarization process as consisting of *interpretation*, *transformation* and *generation*, and in that framework most extractive summarizers can be thought of as only engaging in the first step of interpreting the source document, though extraction itself could perhaps be considered a much simplified transformation stage. The research described in this chapter lays groundwork for the second two steps by exploring properties of spontaneous speech conversations that may aid summarization of a more abstractive variety.

One characteristic of *abstractive* summaries of meetings is that they are normally written from a fairly detached perspective, describing the meeting discussions either from an outsider's perspective or in a manner that synthesizes the important points of the discussion into a form easily understood by a third party. These can be deemed "high-level" summaries because they give a general and broad view of what transpired in the discussion. In contrast, extractive summaries of meetings are comprised of statements actually taken from the meeting discussions themselves. Because of this, the summaries naturally lack the same level of perspective that abstracts have; they indi-

cate what was being discussed at a particular point in time and may be quite specific or technical. In contrast to abstracts, then, they are "low-level" summaries. Among the drawbacks of these low-level extracts are a loss of coherence when a low-level unit is removed from its original context, and a general lack of information richness. It might take two dozen dialogue acts to express the same information as one or two high-level abstract sentences.

To give an example of these differences, the general abstract for one AMI meeting (ES2008a) contains the sentence "The team members discussed the option of combining remotes and how to produce a remote which is capable of controlling multiple devices." There are twelve dialogue acts in the meeting linked to that single abstract sentence, one of which is the statement "If it's just a television that - it's a bit simpler." A person familiar with the topic or with the group might easily be able to surmise what was being discussed at that point, but it is not entirely clear just from reading that dialogue act in isolation. In contrast, if we extracted the dialogue act wherein the project manager says "Like your question earlier, whether this is going to be for television, video or etcetera. Just for television. That's what we're focused on" then the meaning is much more clear, and more information is contained in the latter dialogue act than in the former. Specifically, the project manager has referred to low-level issues in a high-level manner, by explicitly referring to the discussion and the subsequent decision that was made.

This work examines how dialogue acts in spontaneous spoken conversations in the meetings domain vary between low-level and high-level comments. And we are specifically interested in detecting what we call "meta" dialogue acts, where the dialogue acts are not simply high-level in terms of referring to the discussion flow but are also informative in that they synthesize relevant discussion points in a more high-level manner. We also examine how meta dialogue acts can be informative in differing ways - for example, some relate to decisions that have been made while others concern work that remains to be done. The advantages of classifying dialogue acts into classes of meta and non-meta comments is that we can create summaries which are more abstractive in their perspective. This is desirable for two reasons: the dialogue acts comprising such summaries are likely to be more coherent when extracted from their original contexts and concatenated with other informative dialogue acts, and they are also more likely to lend themselves to further interpretation and transformation so that we can ultimately form abstracts more analogous to human abstracts.

7.2 Experimental Setup

This section describes the annotation scheme used for classifying dialogue acts as "meta" or not, and gives a general overview of the experimental setup.

7.2.1 Annotation

One set of annotations that could be to relevant to detecting meta comments in meeting speech are the *reflexivity* annotations for the AMI corpus. A dialogue act is considered to be reflexive if it refers to the meeting or discussion itself. Whereas some dialogue acts refer to the task at hand, such as determining the interface for the remote control, other dialogue acts feature content about how the group *approaches* the task. However, on closer inspection, the reflexivity annotation proves to be insufficient and less than ideal on its own. Many of the dialogue acts deemed to be reflexive consist of statements like "Next slide, please." and "Can I ask a question?" in addition to many short feedback statements such as "Yeah" and "Okay." These are not particularly informative or interesting, despite referring to the flow of discussion at a high level. They refer to trivial aspects of the conversation rather than general overviews of the discussion content. We are not interested in identifying dialogue acts that are *purely* about the flow of discussion, but rather we want to detect dialogue acts that refer to low-level issues in a high-level way. For example, we would find the dialogue act "We decided on a red remote control" more interesting than the dialogue act "Let's move on".

In light of these considerations, we created a novel labelling scheme for meta dialogue acts, using several sources of existing annotation together in order to form a new binary meta/non-meta annotation for the corpus. We now consider a dialogue act to be a "meta" dialogue act if it meets at least one of the following conditions:

- **It is labelled as both extractive and reflexive.**
- **It is labelled as extractive and links to the "decisions" portion of the abstract.**
- **It is labelled as extractive and links to the "actions" portion of the abstract.**
- **It is labelled as extractive and links to the "problems" portion of the abstract.**

The first condition is the ideal class, but it does not occur often in the training data, perhaps four or five such dialogue acts per meeting on average. The remaining conditions use the annotation that links extractive dialogue acts to abstract sentences, as described in Chapter 3 Section 3.3.2 (page 28). The human abstracts are divided into four sections: a general abstract summary, and then sections labelled "decisions", "actions", and "problems." The "decisions" section relates any decisions that were made in the meeting. The "actions" section relates actions that were set out in the meeting such as specific work to do before the next meeting, and the "problems" section details any problems that the group encountered in the meeting. The intuition behind using the dialogue act links to those three abstract subsections is that areas of a discussion that relate to these categories will tend to indicate where the discussion moves from a lower level to a higher level. For example, the group might discuss technical issues in some detail and then make a decision regarding those issues, or set out a course of action for the next meetings. We believe that there is enough commonality between these conditions that they can form a coherent class together, though it does make for noisy training data since we are conflating several sets of annotations.

Since dialogue acts related to decisions, actions and problems are based on links to particular sections of the human abstract, it is worth reviewing the instructions given to the human annotator when writing the abstract. The instructions for the "decisions", "problems" and "actions" sections are as follows:

- **Decisions** Name all decisions that were made during the meeting. Please note that only task-oriented decisions should be included, e.g. "The remote is going to be yellow", while meta-decisions, like "The program manager decided to listen to the designer's opinion first", should not be considered. You can write this section in fragmented text, instead of forming a coherent paragraph.

- **Issues/Problems** Name the problems or difficulties that occurred during the meeting. All problems and/or questions that came to the surface and remained open should be noted in this slot. So should issues that the group managed to solve, if it seems that an amount of time and effort was needed to deal with them. You can also write this section in fragmented text, instead of forming a coherent paragraph.

- **Actions** Name the next steps that each member of the group will take until the next meeting. You can write this section in fragmented text, instead of forming a coherent paragraph.

Note that when annotators are labelling decision dialogue acts, they are instructed not to include meta-decisions such as "The program manager decided to listen to the designer's opinion first". This does not present a problem for our research, and in fact that stipulation is desirable for us. We are not interested in decisions that are purely about the flow of the meeting, but rather decisions concerning the group task.

This work focuses solely on the AMI data, for two reasons: the ICSI data does not contain the "reflexivity" annotation, and the ICSI abstracts have slightly different subsections than the AMI abstracts. Meta dialogue acts constitute less than four percent of the total dialogue acts in the AMI training set; this is in contrast to the work in Chapter 5, where the positive class represented nearly 15% of the total dialogue acts.

7.2.2 Supplementary Features

The experiments described in this chapter use the same features database as used in Chapter 5 (page 68), but we also add two more lexical features that are hypothesized to be of use for this classification task. The first new feature is the number of filled pauses in each dialogue act. This is included because the fluency of speech might change at areas of conversational transition, perhaps including more filled pauses than on average. These filled pauses consist of terms such as "uh", "um", "erm", "mm," and "hmm."

The second new feature is the presence of abstractive or meta cuewords, as automatically derived from the training data. In Chapter 4 on term-weighting, we investigated the usefulness of cuewords for summarization (Section 4.2, page 57). This current chapter explores a more specific type of cueword, and uses the presence of these cuewords as a feature in a machine-learning framework. Since we are trying to create summaries that are somehow more abstract-like, i.e. more high-level, we examine terms that occur often in the abstracts of meetings but less often in the *extracts* of meetings. We score each word according to the ratio of these two frequencies,

$$TF(t,j)/TF(t,k)$$

where $TF(t,j)$ is the frequency of term t in the set of abstracts j from the training set meetings and $TF(t.k)$ is the frequency of term t in the set of extracts k from the training set meetings (see Chapter 4 Section 4.1.1, page 39, for the function definition). This ratio is multiplied by the term's frequency in the training data abstracts so as to avoid small sample sizes. These scores are used to rank the words from most

Feature ID	Description
ENMN	mean energy
F0MN	mean F0
ENMX	max energy
F0MX	max F0
F0SD	F0 stdev.
MPOS	meeting position
TPOS	turn position
DDUR	d. act duration
PPAU	precedent pause
SPAU	subsequent pause
UINT	uninterrupted length
WCNT	number of words
DOMD	dominance (d. acts)
DOMT	dominance (seconds)
ROS	rate of speech
SUI	su.idf sum
TFI	tf.idf sum
ACUE	abstractive cuewords
FPAU	filled pauses

Table 7.1: Features Key

abstractive to least abstractive, and we keep the top 50 words as our list of high-level terms. Appendix B (page 179) lists the top 50 cue terms derived from the training data. The top 5 abstractive cuewords are "team", "group", "specialist", "member", and "manager" (these represent stems, and so "group" will match "groups" and "grouped", etc.). Unlike the work described in Chapter 4 Section 4.2, where we began with a list of cuewords that were hypothesized to be informative, these cuewords are learned entirely from the data. As a consequence, the list of terms is somewhat noisier and also contains a few terms that are specific to the domain of the AMI meetings. For example, one word on the list of top cuewords is "remote." It may at first seem surprising that this word would occur much more often in abstracts than in extracts, but it is most likely due to the fact that in meetings participants will often refer to the remote using pronouns or otherwise refer indirectly. The vast majority of the abstractive cuewords are not specific to this corpus.

For both the manual and ASR feature databases, each dialogue act then has a feature indicating how many of these high-level terms it contains.

Table 7.1 lists the 19 features and their IDs for ease of reference.

7.2.3 Summarization Experiments

We again build a logistic regression classifier by training on the AMI training data, as in Chapter 5, this time incorporating 19 features instead of 17. Feature subset selection is carried out as before. The classifier output for the test set is used to create summaries of 700 words length, which we evaluate using two separate instantiations of the weighted precision/recall/f-score method described earlier. The classifier itself is evaluated using the ROC and AUROC measures.

7.3 Evaluation

We use several types of evaluation for these new summaries: two implementations of weighted precision/recall/f-score based on new and old extractive labelling schemes, as well as evaluation using ROUGE, which does n-gram comparison between machine summaries and reference gold-standard summaries.

7.3.1 Weighted Precision With New Extractive Labels

The new weighted precision/recall/f-score evaluation is the same as the old method but simply using new labels based on the criteria for extraction described above. So, many dialogue acts which were previously considered extractive are now considered non-extractive. The positive class is a very small subset of the original positive class. This evaluation measures how much these relatively short summaries incorporate dialogue acts related to decisions, actions, problems and reflexivity.

7.3.2 Weighted Precision With Old Extractive Labels

We also evaluate the summaries using the original formulation of weighted precision/recall/f-score, with the previous extractive/non-extractive labels, simply for comparison with the results of previous chapters. It is not expected that the new meta summaries will fare as well using the original formulation of the metric, since the vast majority of extract-worthy dialogue acts are now considered members of the negative class and the evaluation metric is based on the previous extractive/non-extractive labels, but the results are included out of interest nonetheless.

7.3.3 ROUGE

Our research until now has not utilized the ROUGE metric, partly because of negative correlations with human judgements found in previous work (Murray et al., 2005b, 2006), and also because the multiple human extracts already provide sufficient gold-standard evaluations, but we use ROUGE evaluation in this chapter because it seems very applicable to the task at hand. We are aiming to create extracts that are more abstract-like, and ROUGE compares a machine summary to multiple human abstracts. It is hypothesized that ROUGE will be sensitive to the differences between this new type of summary and the summaries created previously that were based purely on informativeness rather than perspective.

We use ROUGE-1.5.5 and focus on the ROUGE-2 and ROUGE-SU4 metrics, as described in Chapter 3 Section 3.5.4.2 (page 34), which have previously been found to correlate well with human judgements in the DUC summarization tasks (Lin, 2004; Dang, 2005). We calculate precision, recall and f-score for each, and ROUGE is run using the parameters utilized in the DUC conferences, plus removal of stopwords:

ROUGE-1.5.5.pl -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d -s -a

These parameters indicate that ROUGE-2 and ROUGE-SU4 are to be calculated, with stemming and removal of stopwords, that precision and recall are weighted equally, and that the confidence interval is 95%.

For each meeting in the AMI test set, there are two gold-standard human abstracts used for comparison. Ideally we would like several reference summaries per meeting, but additional human abstracts have yet to be created for this corpus.

7.4 Results

In this section we provide an overview of the various evaluation results and a detailed analysis of the features used.

7.4.1 Classification Results

When training on the manual transcript aligned database, the optimal feature subset is 13 features, which excludes mean F0, position in the speaker's turn, precedent pause, both dominance features, and filled pauses. The best five features in order are *su.idf*, dialogue act word-count, *tf.idf*, dialogue act duration, and uninterrupted duration. Whereas in Chapter 5 Section 5.6.1 (page 78) we found that the optimal subset

for summarization of manual transcripts was 17 features, we find here that there are fewer features that are useful for discerning high-level informative dialogue acts from dialogue acts that are either uninformative or informative but low-level.

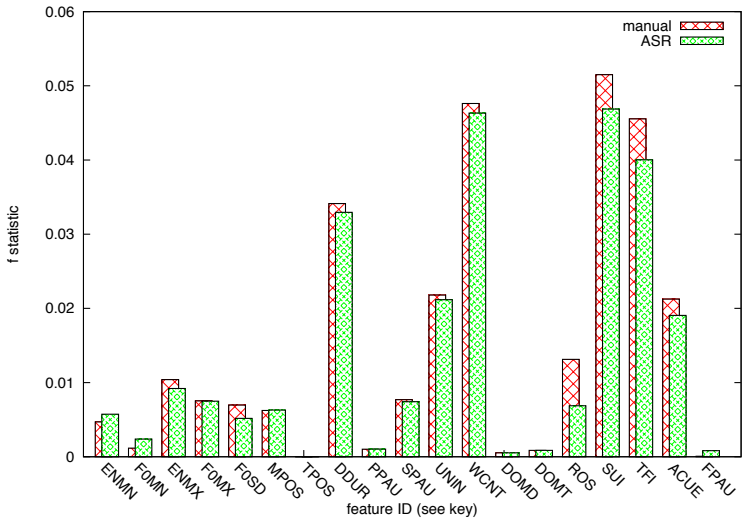


Figure 7.1: f statistics for AMI database features

When training on the ASR aligned database, the optimal feature subset is the entirety of the 19 features. The best five features in order are *su.idf*, word-count, *tf.idf*, dialogue act duration, and uninterrupted duration, the same as for manual transcripts.

Figure 7.1 gives the histograms for the feature f statistics using both the manual and ASR databases.

The ROC curves for the logistic regression classifier applied to the 20-meeting test set are shown in Figure 7.2, for manual and ASR. For manual, the AUROC is 0.843. For ASR, the AUROC is 0.842. Chance level classification would exhibit an AUROC of around 0.5, represented by a diagonal ROC curve from (1,1) to (0,0) as the posterior probability threshold increases.

This result is very encouraging, as it shows that the classifier can discriminate between high-level informative dialogue acts on the one hand, and dialogue acts that are either uninformative or are informative but low-level on the other hand. Given that we created a new positive class based on whether or not a dialogue act satisfies one of four criteria, and that we consider everything else as negative, this result shows that dialogue acts that meet at least one of these extraction criteria do have characteristics in common with one another and can be discerned as a separate group from the remainder.

Appendix E (page 187) provides sample meta and non-meta summary output for

AMI meeting TS3003c.

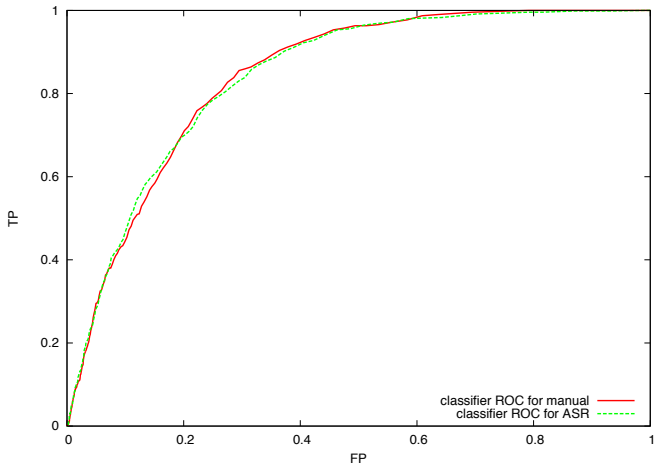


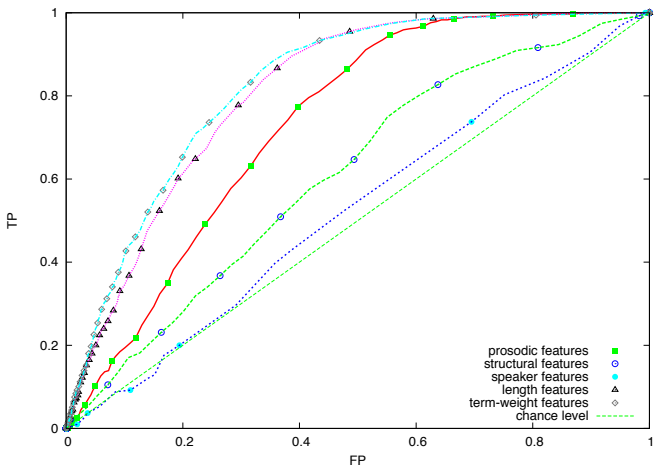
Figure 7.2: ROC Curves for LR Classifiers on AMI data

7.4.2 Features Analysis

As with Chapter 5 Section 5.6.1.1 (page 80), we could like to know how various feature subsets can perform relative to the full feature set classification. We already know from Section 7.4.1 that the optimal sets are 13 and 19 features for manual and ASR, respectively, but it is nonetheless interesting to inspect how certain classes of features contribute to classification performance and how well they do on their own.

The feature classes are the same as in Chapter 5, but with the two additional features of filled pauses and abstractive cuewords under the *lexical* category along with the previously-used term-weight features:

- **Prosodic features**: The features of energy, pitch, pause, and rate-of-speech, for a total of 8 features.
- **Length features**: The features of total dialogue act length, uninterrupted length, and dialogue act duration, for a total of 3 features.
- **Speaker features**: The two features of speaker dominance are considered as a class of their own.
- **Structural features**: There are two structural features: the position of the dialogue act in the meeting and the position in the speaker's turn .
- **Lexical features**: Abstractive cuewords, filled pauses, *tf.idf* and *su.idf*.



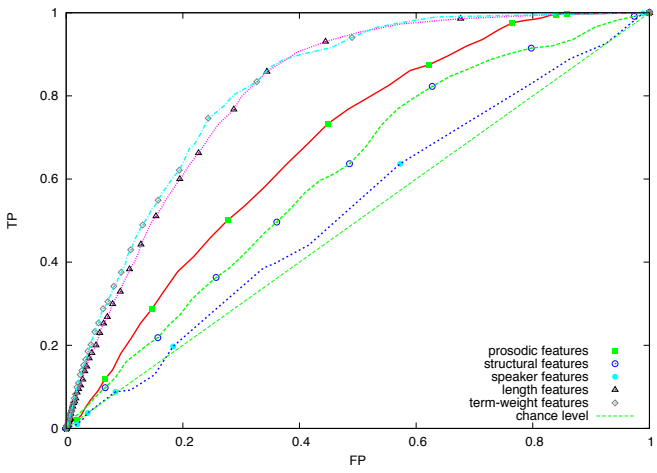
Fea. Subset	AUROC
Prosodic	0.734
Structural	0.611
Speaker	0.524
Length	0.811
Term-Weight	0.826

Table 7.2: AUROC Values, Manual Transcripts

Figure 7.2 shows the ROC curves for each feature subset for the manual database. The AUROC for prosodic features is 0.734, for speaker features is is 0.524, for structural features it is 0.611, for length features it is 0.811 and for term-weight features the AUROC is 0.826. We find that no individual subset matches the classification performance found by using the entire feature set, but that several classes exhibit credible individual performance. The length and term-weight features are clearly the best, but again we find that prosodic features alone can discern these classes to a respectable degree.

Figure 7.3 shows the ROC curves for each feature subset for the ASR database. The AUROC for prosodic features is 0.67, for speaker features is is 0.55, for structural features it is 0.632, for length features it is 0.811 and for term-weight features the AUROC is 0.820. The trend is largely the same as above: no individual feature type is better than the combination of feature types. The principal difference is that prosodic features alone are worse on ASR and term-weight features are about the same as on manual. A similar finding was reported in Chapter 5 Section 5.6.1.1 (page 80). It seems counter-intuitive perhaps that prosodic features are slightly worse and term-weight features are the same or slightly better on noisy ASR data, but the prosodic features depend on word segmentation and so can degrade when there are ASR errors. For example, insertions might lead to taking F0 readings where there are no words, resulting in skewed F0 ranges.

The meta dialogue acts can be characterized as having higher mean energy and pitch levels, much higher maximum energy and pitch levels, and higher pitch standard



Fea. Subset	AUROC
Prosodic	0.684
Structural	0.612
Speaker	0.527
Length	0.812
Term-Weight	0.822

Table 7.3: AUROC Values, ASR Transcripts

deviation. They tend to occur later in the meetings, a finding that is the opposite of findings in Chapter 5, wherein generally informative dialogue acts are slightly more likely to occur early on in a meeting. They are on average twice as long in duration as other dialogue acts, with much longer precedent pauses. There is less likely to be subsequent pause - in fact, there tends to be speaker overlap at the end of meta dialogue acts. Thus, the uninterrupted duration is much shorter on average than the full duration, but still more than twice as long on average than non-meta dialogue acts. The dialogue acts are much longer in terms of words, averaging nearly 14 per dialogue act. They are more likely to be uttered by the dominant speaker in the meeting, according to both dominance criteria. The rate-of-speech is much faster than for the negative class. For both term-weighting criteria, the positive class scores much higher on average - nearly three times as high as the negative class. Meta dialogue acts are much more likely to have abstract cuewords, but only slightly more likely to have filled pauses.

7.4.3 Evaluating Summaries

Table 7.4 presents the weighted f-scores using the novel extractive labelling, for the new meta summaries as well as for the summaries created and evaluated in Chapter 5. For manual transcripts, the new summaries outperform the old summaries with an average f-score of 0.17 versus 0.12. The reason for the scores overall being lower than the f-scores reported in Chapter 5 using the original formulation of weighted precision/recall/f-score is that there are now far fewer positive instances in each meeting since we are restricting the positive class to the “meta” subset of informative dia-

Meet	NonMeta-Man	NonMeta-ASR	Meta-Man	Meta-ASR
ES2004a	0.08	0.07	0.10	0.11
ES2004b	0.04	0.0	0.04	0.06
ES2004c	0.11	0.1	0.22	0.18
ES2004d	0.24	0.18	0.10	0.20
ES2014a	0.19	0.28	0.22	0.31
ES2014b	0.04	0.01	0.10	0.09
ES2014c	0.09	0.12	0.16	0.15
ES2014d	0.14	0.19	0.17	0.17
IS1009a	0.21	0.28	0.35	0.36
IS1009b	0.04	0.06	0.10	0.13
IS1009c	0.0	0.0	0.22	0.21
IS1009d	0.14	0.23	0.11	0.08
TS3003a	0.26	0.22	0.27	0.23
TS3003b	0.10	0.12	0.18	0.25
TS3003c	0.03	0.18	0.28	0.30
TS3003d	0.12	0.18	0.14	0.14
TS3007a	0.33	0.25	0.36	0.40
TS3007b	0.05	0.05	0.09	0.08
TS3007c	0.13	0.15	0.14	0.22
TS3007d	0.14	0.15	0.11	0.14
AVERAGE	**0.12**	**0.14**	**0.17**	**0.19**

Table 7.4: New Weighted F-Scores on Manual and ASR Transcripts for Meta and Non-Meta Approaches

Meet	Meta-Man	Meta-ASR
ES2004a	0.39	0.41
ES2004b	0.15	0.15
ES2004c	0.21	0.18
ES2004d	0.12	0.17
ES2014a	0.36	0.47
ES2014b	0.16	0.19
ES2014c	0.15	0.14
ES2014d	0.12	0.15
IS1009a	0.37	0.39
IS1009b	0.14	0.14
IS1009c	0.23	0.23
IS1009d	0.21	0.18
TS3003a	0.44	0.40
TS3003b	0.20	0.22
TS3003c	0.28	0.28
TS3003d	0.18	0.18
TS3007a	0.39	0.43
TS3007b	0.15	0.14
TS3007c	0.13	0.17
TS3007d	0.11	0.14
AVERAGE	**0.23**	**0.24**

Table 7.5: Old Weighted F-Scores on Manual and ASR Transcripts for Meta Approaches

logue acts. The meta summaries are significantly better than the previous summaries on this evaluation according to paired t-test ($p<0.05$).

For ASR, we find both the new meta summaries and older non-meta summaries performing slightly better than on manual transcripts according to this evaluation. The meta summaries again are rated higher than the non-meta summaries, with an average f-score of 0.19 versus 0.14. The meta summaries are again significantly better than the previous summaries according to paired t-test ($p<0.05$).

Table 7.5 presents the weighted f-scores for the new meta summaries using the original formulation of weighted precision/recall/f-score, where the classes are general informativeness versus uninformativeness. As mentioned earlier, it would not be expected that the new meta summaries would compare well with previous summaries according to this metric, as the evaluation is based on the original extractive/non-extractive classes and the new summaries are based on a severely restricted subset of this positive class, with the remainder considered negative instances. Quite surprisingly, the weighted f-scores for these new summaries are actually slightly higher than the f-scores reported in Chapter 5 (page 81). The f-score for manual transcripts is 0.23 compared with 0.21 previously, and 0.24 for ASR compared with 0.22 earlier. This

is quite a surprising and encouraging result, that our new annotation and subsequent "meta" machine-learning experiments have led not only to improved general informativeness but also to finding areas of high-level meta comments in the meetings.

To determine why these weighted f-score results are unexpectedly higher, we calculate the annotator kappa statistic for the current annotation scheme, where a dialogue act is considered positive if it satisfies one of the four stated criteria and considered negative otherwise, and we also calculate the kappa statistic for the case where a dialogue act is positive if it is linked to the general *abstract* section of the human summary and considered negative otherwise. The kappa statistic is a way of evaluating how closely two annotators agree with each other on an annotation task. The statistic is derived by calculating

(Observed Agreement - Chance Agreement) / (1 - Chance Agreement)

For each meeting in the corpus, the kappa value for each annotator pair is calculated and these values are averaged to derive a single kappa value for that meeting. These averages are then summed and averaged over the corpus to derive an average kappa statistic for the corpus. For the annotation scheme presented here based on four related criteria, the annotator agreement is **0.45**, whereas the kappa statistic for dialogue acts considered generally informative or uninformative is **0.40**. While these numbers in general are somewhat low, reflecting the difficulty of summarization annotation, the results show that it is substantially easier for annotators to agree upon informativeness when there is a specific criterion on which to rate a dialogue act, compared with simply stating that a dialogue act is informative or not. This difference in kappa statistics is apparently the reason why the new meta summaries perform slightly better than the previous summaries which otherwise would be expected to be more generally informative; annotator agreement is simply higher on that data.

7.4.3.1 ROUGE Results

In this section we present the ROUGE results for our new meta summaries in comparison with our previously generated summaries. We also compare the performance of these automatic extracts to human extracts of the same length.

The ROUGE results are very encouraging for our new meta summaries, according to both the ROUGE-2 and ROUGE-SU4 metrics, with the new summaries outperforming the summaries described in Chapter 5.

For ROUGE-2, using manual transcripts, the meta summaries average a score of 0.039, compared with 0.033 for the previous non-meta summaries, a significant im-

Meet	NonMeta-Man	NonMeta-ASR	Meta-Man	Meta-ASR	Human
ES2004a	0.02548	0.03020	0.02675	0.02922	0.05105
ES2004b	0.00537	0.00801	0.01894	0.02070	0.01735
ES2004c	0.03596	0.02086	0.05107	0.04432	0.02675
ES2004d	0.01618	0.01396	0.02216	0.01108	0.01362
ES2014a	0.03283	0.03380	0.02323	0.05690	0.08037
ES2014b	0.01496	0.03133	0.01326	0.02696	0.03518
ES2014c	0.02656	0.04307	0.04284	0.02906	0.07938
ES2014d	0.04961	0.04897	0.05645	0.04762	0.06133
IS1009a	0.09370	0.05191	0.07244	0.06557	0.14720
IS1009b	0.02213	0.01565	0.02449	0.01233	0.06278
IS1009c	0.02401	0.02620	0.06076	0.06470	0.10256
IS1009d	0.06793	0.06667	0.04959	0.04829	0.08995
TS3003a	0.04878	0.04408	0.0508	0.04348	0.04558
TS3003b	0.03250	0.02944	0.02762	0.06096	0.04234
TS3003c	0.01530	0.02076	0.08078	0.08174	0.06541
TS3003d	0.04218	0.04404	0.03816	0.04272	0.05155
TS3007a	0.05053	0.04188	0.05676	0.06316	0.08254
TS3007b	0.01658	0.03000	0.01395	0.01750	0.01591
TS3007c	0.02693	0.01840	0.02478	0.02982	0.06069
TS3007d	0.01385	0.02228	0.01785	0.02266	0.02417
AVERAGE	**0.033**	**0.032**	**0.039**	**0.041**	**0.058**

Table 7.6: ROUGE-2 Scores on Manual and ASR Transcripts for Meta and Non-Meta Approaches

provement ($p<0.1$). On ASR transcripts, the meta summaries score slightly higher with an average of 0.041 compared with 0.032 for the non-meta summaries, another significant result ($p<0.05$). Table 7.6 gives the results for each summary type on both manual and ASR transcripts.

According to ROUGE-SU4, on manual transcripts the meta summaries outperform the low-level summaries with an average of 0.066 compared with 0.061, respectively. On ASR transcripts, the meta summaries average 0.069 compared with 0.064 for the low-level summaries. For each transcript type, the rating of the meta summaries is significantly better than that of the low-level summaries according to ROUGE-SU4 (both $p<0.05$). Table 7.7 gives the ROUGE-SU4 results for both summary types on manual and ASR transcripts.

On average, the human summaries are still considerably superior to the best automatic extracts of the AMI test set meetings, according to both ROUGE-2 and ROUGE-SU4, but there are several meetings for which the automatic summarizers approach or exceed human-level performance. As discussed in Chapters 4 and 5, it is more difficult to attain human-level performance on the AMI data versus the ICSI data, and so there

Meet	NonMeta-Man	NonMeta-ASR	Meta-Man	Meta-ASR	Human
ES2004a	0.04292	0.04694	0.04284	0.04431	0.06016
ES2004b	0.04424	0.04670	0.04825	0.05132	0.05664
ES2004c	0.05537	0.05519	0.07447	0.08006	0.07204
ES2004d	0.04714	0.05310	0.04847	0.04520	0.04812
ES2014a	0.06719	0.06518	0.05374	0.08664	0.10463
ES2014b	0.06957	0.07419	0.06200	0.06800	0.07813
ES2014c	0.06521	0.09326	0.07834	0.07019	0.10030
ES2014d	0.06936	0.07539	0.07911	0.07161	0.07982
IS1009a	0.10793	0.08946	0.08860	0.09835	0.15256
IS1009b	0.05677	0.05430	0.06362	0.06331	0.08556
IS1009c	0.04680	0.04851	0.07309	0.06862	0.11776
IS1009d	0.09827	0.09860	0.09778	0.08030	0.12989
TS3003a	0.05788	0.06348	0.06892	0.06732	0.05704
TS3003b	0.06729	0.06809	0.05992	0.09610	0.07534
TS3003c	0.04440	0.06044	0.09543	0.09618	0.10064
TS3003d	0.07215	0.07361	0.07186	0.07348	0.07626
TS3007a	0.06261	0.05370	0.07364	0.07212	0.09572
TS3007b	0.05374	0.06177	0.05547	0.05336	0.05687
TS3007c	0.04443	0.04728	0.04123	0.05178	0.07990
TS3007d	0.03961	0.04405	0.03907	0.04385	0.05595
AVERAGE	**0.061**	**0.064**	**0.066**	**0.069**	**0.084**

Table 7.7: ROUGE-SU4 Scores on Manual and ASR Transcripts for Meta and Non-Meta Approaches

remains room for improvement.

In comparing ROUGE scores between different summarization approaches across different domains, there are several factors to consider. One is that ROUGE was originally a recall metric but now calculates recall, precision and f-score. While it is subsequently more common for researchers to report f-scores, many still present only recall measures. Second, there are many different compression rates used by different researchers - some extract until reaching a certain word percentage and others extract a certain percentage of dialogue acts, among other methods. Comparing the resultant summaries is less difficult now that ROUGE calculates f-score, but with only recall scores it is challenging to compare summaries of varying lengths. Three, some researchers use human abstracts as the reference summaries (e.g. (Murray et al., 2005a)) while others use human extracts (e.g. (Galley, 2006)). While we use human abstracts in this chapter, in Chapter 6 we used human extracts due to the lack of multiple abstracts for those meetings. Four, ROUGE contains many parameters that can be adjusted by the user, and reporting the parameters used is critical to replication and comparison. For example, excluding or including stopwords can change the range of scores dramatically for all systems.

To put our ROUGE scores in context with state-of-the-art summarization systems, in the DUC 2007 main task and pilot task the ROUGE-2 scores ranged from 0.036 to 0.124. Above we reported a best ROUGE-2 average of 0.041. If we run ROUGE again with the exact DUC parameters (i.e. the same parameters as above but now including stopwords) our ROUGE-2 average for the meta system is 0.064. The DUC 2007 ROUGE-SU4 scores range from 0.074 to 0.177. Above we reported a best ROUGE-SU4 average of 0.069. With the DUC parameters, the ROUGE-SU4 average for our meta system is 0.12. It can be seen that we are well within the range of DUC systems' scores, while working with much noisier data. Galley (2006) reported ROUGE-2 recall scores of 0.42-0.44 for the ICSI test set with summary compression set at 12.7% of the word count, and ROUGE-2 f-score of 0.64 when selecting 10% of dialogue acts (much longer than our current summaries). It is difficult to compare scores because of differing compression rates and because Galley used human extracts as reference summaries. However, in Chapter 6 Section 6.4.3 (page 117) we used human extracts as references and achieved ROUGE-2 f-scores averaging 0.55 for summaries that were between 30% and 40% of the total meeting word count. That is also comparable to the highest ROUGE-2 f-score reported by Maskey and Hirschberg (2005), which is 0.544 when selecting 23% of sentences. They also report AUROC scores, with a highest

AUROC of 0.771, compared with our best AUROC here, 0.843.

It is worth noting generally that ROUGE scores are much lower when using short human abstracts as gold-standard references (e.g. DUC system scores and most of our ROUGE scores) compared with using longer human extracts as references summaries (e.g. Galley's results and our results in Chapter 6).

7.4.3.2 Decision Audit Revisited

In Chapter 6 (page 93) we describe an extrinsic evaluation in the form of a *decision audit* task. In that set of experiments, users review several meetings in order to satisfy a complex information need, using different information sources in each experimental condition. For the two extractive conditions – one on manual transcripts and one on ASR – the extracted dialogue acts are just deemed to be generally informative, without any consideration of low-level or high-level perspective as we are addressing in this chapter. However, by revisiting the decision audit results and analyzing user behaviour more closely, we can see how useful the manually-labelled meta dialogue acts are for browsing in that scenario. In other words, we would like to know to what degree the decision audit participants rely on meta dialogue acts in the extractive summaries compared to the other dialogue acts. Specifically, we look at instances of summary clicks, where a participant clicks on a summary dialogue act in order to navigate to that part of the transcript and audio-video record.

For Condition EM, which incorporated extractive summaries of manual transcripts, 38% of extractive summary clicks on average are clicks on meta dialogue acts, as defined in this chapter: dialogue acts that are reflexive, or related to decisions, actions or problems in the meeting. Considering that such dialogue acts comprise only 23% of extracted dialogue acts in total, it is surprising and encouraging that they represent such a high proportion of summary clicks.

For Condition EA, on ASR transcripts, the story is similar. In this condition, 31% of extractive summary clicks are on meta dialogue acts on average, despite these dialogue acts comprising only 20% of the total dialogue acts extracted for that condition.

In both cases, participants use meta dialogue acts much more often than would be expected based simply on the frequency of those dialogue acts in the extractive summaries. It is possible that users are more quickly able to understand and process these dialogue acts and therefore are more likely to use them as indices into meeting record. It might also be the case that these dialogue acts more obviously represent significant moments or turning-points in the meeting discussions and therefore are

good candidates for further browsing.

These patterns from the decision audit data give a good deal of extrinsic motivation for researching the differences between meta and non-meta dialogue acts. Users clearly find such summary units to be informative and useful beyond our prior expectation.

7.5 Discussion

There are several compelling results to the experiments presented herein. First of all, we find that our combination of annotations into a single "meta" annotation was successful in terms of identifying a group of dialogue acts that share common characteristics and can be discerned from the remainder of the dialogue acts. These dialogue acts are also realized quite differently from the generally informative dialogue acts, in terms of prosodic and structural characteristics. According to the ROC and AUROC evaluations, the classifiers perform very well in terms of the true-positive/false-positive ratio.

We also find that the optimal classification according to AUROC utilizes a variety of features, showing that it is advantageous to explore characteristics beyond lexical features, incorporating prosodic and structural cues. Even classification using prosodic features alone results in decent classification performance. The optimal feature subset according to balanced accuracy is 13 features for manual transcripts and the entirety of the feature set for ASR transcripts. Of the two new features, abstractive cuewords are very useful but filled pauses are less useful and are excluded during feature subset selection on manual transcripts. It may be the case that filled pauses are useful for more generally classifying dialogue acts as informative or uninformative but simply less useful for discerning meta dialogue acts from the negative class.

Compared with the results in Chapter 5, in creating these novel meta summaries it is even more imperative to use a variety of multi-modal features in order to achieve optimal results. In that previous chapter, we found that using a combination of all features was consistently the best approach but that length and term-weight features somtimes performed competitively on the test sets. Here we find that the best feature type subset performs substantially worse than the combination of feature types. For example, term-weight features for manual transcripts have an AUROC of 0.826 compared with 0.843 for the selected feature combination, and for ASR they result in an AUROC of 0.822 compared with 0.842 for the full feature set.

For the brief 700-word summaries we generated, we find that the new meta sum-

maries are significantly better in terms of the new weighted f-score metric, and slightly better in terms of the original weighted f-score metric. The latter result is very much a surprise, as that evaluation metric is based on annotations of general informativeness or uninformativeness, and it would have been expected that training on the original annotation would therefore be superior. The reason seems to be that annotator agreement is higher when classifying these high-level dialogue acts compared with labelling dialogue acts as just generally informative or not.

The ROUGE results are also very encouraging, with the new meta summaries outperforming the low-level summaries according to both the ROUGE-2 and ROUGE-SU4 metrics, the most generally reliable of the ROUGE suite of metrics. These findings lend evidence to back up our intuition that the new meta summaries are more similar to human abstracts.

By referring back to the decision audit evaluation and more closely analyzing users' browsing behaviour in the extractive conditions, we are able to show that participants utilize the meta dialogue acts much more than would be expected based on how frequent such dialogue acts occur in the summaries. This provides evidence that our novel research on meta dialogue act classification is justified from an extrinsic point of view; users find these meta dialogue acts to be useful indices to the meeting record.

There are interesting similarities between this work and other types of dialogue act classification such as *decision detection* (Hsueh et al., 2007). The output of a decision classifier can be thought of as a focused extractive summary, locating dialogue acts that are informative for a particular reason. Like our current work, such classification approaches are capable of creating more intelligent extractive summaries by looking beyond the simple distinction of informative versus uninformative and instead basing extraction on more specific relevance criteria.

There is also some comparison between this work and the work of Teufel and Moens (1999). In creating abstracts of scientific articles, they viewed the abstract as a template with slots relating to rhetorical roles such as *background*, *purpose*, *solutions*, and *conclusions*. In a first supervised classification step, they attempted to extract sentences that related to any of those rhetorical roles, and in a second classification step they tried to assign the correct rhetorical role to each extracted sentence. Their expressed desire was to create automatic summaries that were more than "just a collection of sentences." Like our current work, the extraction criteria are more meaningful and the output is correspondingly more flexible. In addition to features such as location, title overlap and sentence length, the authors also incorporated meta comments for the

articles domain, such as “we have argued...”. While a meta perspective is a much more common feature of meeting speech than most formal text data, it was useful for their domain and would possibly be useful for other text data. For example, newswire articles sometimes contain phrases such as “it has been reported...” and “early reports are that...” and these could potentially be exploited for summarization in the news domain.

7.6 Conclusion

This research has taken some first steps in creating speech summaries that do more than merely indicate which dialogue acts are informative. We are able to classify dialogue acts which relate to high-level aspects of the meeting discussion, including dialogue acts relating to decisions, actions and problems encountered in the discussion. We hypothesize that not only do such dialogue acts provide a better perspective of the meeting discussion than low-level technical dialogue acts, but that users would find them easier to understand when removed from their original contexts. This can be tested by future extrinsic evaluations, and we have gone some way toward proving this hypothesis here by revisiting the decision audit results.

We have also shown that annotators have more difficulty in agreeing on whether something is simply informative or uninformative, compared with annotating dialogue acts that are informative for a particular reason. Not only is annotator agreement higher in the latter cases, but a summarizer that can classify a given dialogue act as being informative for a particular reason is much more flexible in terms of creating a variety of final summary structures. For example, by training on individual classes one could create a summary that first lists dialogue acts relating to decisions, followed by dialogue acts that identify action items for the following meeting. A hierarchical summary could also be created, with high-level dialogue acts at the top, linked to related lower-level dialogue acts that might provide more detail.

Finally, it is hypothesized that such meta summaries will be useful for moving summarization research further down the extractive-abstractive continuum, by lending themselves to further transformations and the generation of novel sentences about the meeting content.

Chapter 8

Further Work

In this chapter we introduce several issues that stem from the current work and discuss how they might be addressed in the near future: automatic compression, online summarization, and spurt-based summarization. We first characterize each of the issues, discuss some possibilities for addressing them, and present results of preliminary responses to these challenges. We then conclude by discussing future directions for this work, and briefly describe other challenges within the field.

8.1 Dialogue Act Compression

8.1.1 Introduction

In automatic speech summarization systems, it has been shown that the length of an utterance or dialogue act in seconds or in number of words is a very helpful feature in determining informativeness for inclusion in an extractive summary (Maskey & Hirschberg, 2005; Murray et al., 2006). This has been attested throughout this thesis, that features related to dialogue act length are very indicative of informativeness. Consequently, summarizer output will likely consist of a concatenation of lengthy dialogue acts. If the compression rate for summarizing an hour-long meeting is quite low, then few dialogue acts will be extracted. For that reason, it is very desirable to automatically compress these dialogue acts so that more can be extracted without exceeding the overall length limit.

Many automatic sentence compression techniques rely on deriving the syntactic parse of a given sentence (Knight & Marcu, 2000; McDonald, 2006). Because spontaneous speech tends to be very fragmented and disfluent, successful parsing of speech

often relies on the detection and correction of disfluencies wherever possible (Nakatani & Hirschberg, 1993; Charniak & Johnson, 2001). Errorful ASR transcripts of disfluent speech add another layer of difficulty for syntactic parsers. Here we investigate whether we can avoid the sequence of challenges inherent in disfluency correction and speech parsing by carrying out compression without any syntactic information. This section therefore explores the use of prosody in compressing informative dialogue acts from meeting speech. More specifically, the techniques described below compress the dialogue acts by trying to preserve the original pitch contour as much as possible in the compressed dialogue act. The simple intuition behind this method is that prosody is reflective of meaning (Steedman, 2000, 2007) and that preserving this aspect of the prosody may preserve a great deal of the meaning as well.

Two methods of using prosody for speech compression are described below. They are first evaluated subjectively by humans grading on both informativeness and readability criteria, alongside human-authored gold-standards and random baseline compressions. The second evaluation is edit distance, objectively measuring the string distance between the automatic approaches and the gold-standards. In addition to the prosodic and random approaches, a simple text compression method is implemented and included for this edit distance evaluation.

8.1.2 Previous Work

In work by Hori et al. (2002), T. Kikuchi and Hori (2003), a sentence compression method is described and results on English and Japanese broadcast news are given. The authors use word confidence scores, word significance scores, trigram language scores, and word concatenation scores to determine the optimal compression of a given sentence. The difference between the language score and the word concatenation score is that the former relies solely on trigram language probabilities while the latter is based on the dependency structure of the sentence. For example, a given compression may have a high language score but violate the dependency structure of the original. The dynamic programming method for finding the optimal compression is described in Hori and Furui (2004).

Again on Broadcast News data, Kolluru et al. (2005) present a multi-stage compaction method using a sequence of multi-layer perceptrons. First, confidence scores are used to remove incorrectly transcribed words. A chunk parser identifies intra-sentential chunks and a subset of the chunks are then chosen based on the presence of

Named Entities and *tf.idf* scores.

Clarke and Lapata (2006) present a compression method for both text and speech data, using the Ziff-Davis and Broadcast News corpora, respectively. Their scoring function consists of an n-gram language model coupled with several constraints, with the optimal compression given the constraints determined by Integer Programming. The constraints are linguistically motivated and include stipulations such as requiring that a compression contain at least one verb if the original sentence contains at least one verb, and that if a verb is selected for the compression then its arguments are selected as well. Clarke and Lapata also indicate their intention to apply their compression method to meeting data. Reported results are on manual transcripts.

Ohtake et al. (2003) use prosodic features for speech-to-speech newscast compression and therefore do not use ASR at all. They locate accent phrase boundaries by analyzing fall-rise F0 patterns, determine which adjacent accent phrases belong together as single summary units, and then compare two prosodic methods for selecting the most important summary units. For example, summary units can be eliminated if their mean energy level falls below a pre-determined threshold or if a derived F0 summary unit score is above a speaker-dependent threshold. The authors also attempt to use prosodic features to determine whether a given summary unit depends on the preceding summary unit, so that when a summary unit is eliminated, its dependants are also eliminated. Because broadcast news usually presents the most important information first, all summary units from the first sentence are selected.

The approaches described above are all applied to broadcast news speech. Because broadcast news data contain both read and spontaneous portions, the challenges for automatic compression may be slightly different than for meetings, which normally feature purely spontaneous speech. Below we describe our novel compression methods for such meeting data.

8.1.3 Compression Methods

This section presents the compression methods in detail. First, two prosodic methods are described, both of which strive to compress the utterance by preserving the pitch contour. A simple textual method is presented, as well as a baseline compression method. A first step for each method is to remove simple filled pauses such as *uh* and *erm* as well as immediate repetitions of a word. The compression rate is between 0.65 and 0.70 for all of the automatic compression methods.

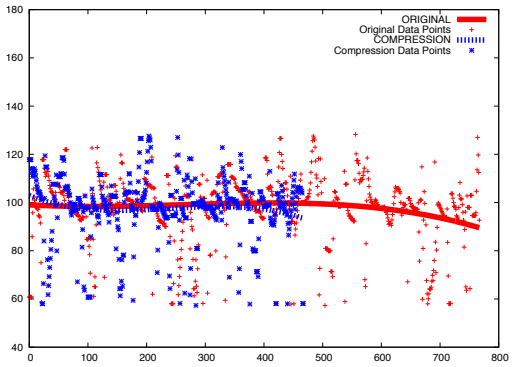


Figure 8.1: *Sample Dialogue Act and Summary Contours (first prosodic method)*

for each *dialogue act*
  try: *break dialogue act into prosodic phrases using pause duration*
  if not: *at least three segments*
    then*: segment using pitch reset*
  for each: *prosodic phrase*
    calculate pitch slope(phrase)
    for each: *word*
      calculate pitch slope(word)
  current length = 0
  while: *current length* $<$ *desired length*
    for each: *prosodic phrase*
      if not: *at least three words*
        then: *skip*
      else: *select* $\arg\min_{word} f(word) = (abs(slope(word) - slope(phrase)))$
      remove selected word from candidate set
      current length += 1

Table 8.1: First Compression Algorithm

8.1.3.1 Prosodic Methods

The first prosody method begins by breaking the utterance into prosodic phrases or chunks. The primary cue for phrase boundary is pause length, with pauses of 100 ms or more being considered a boundary. A secondary method is to look for instances of pitch reset which would signal the beginning of a new prosodic phrase. More specifically, we are looking for areas where the pitch falls to a low level for at least 300 ms before rising sharply again, with the fall-rise pattern signalling the pitch declination of one phrase and the beginning of another. We first attempt to locate the boundaries using only pause, as it is considered more reliable, but if we are unable to break the dialogue act into at least 3 chunks, we revert to looking at pitch reset as well.

Once the prosodic phrases are located, the overall pitch slope for each phrase is measured. We then begin an iterative process, wherein for each phrase we measure

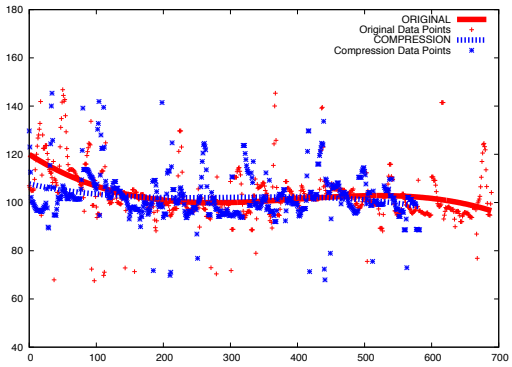


Figure 8.2: *Sample Dialogue Act and Summary Contours (second prosodic method)*

the pitch slopes of its constituent words and select the word whose slope is closest to that of the phrasal slope. If a phrase has no more than two words, we skip it altogether as it is likely to be a disfluent fragment. We continue the iterative process until the desired number of words has been selected for the compression. Table 8.1 gives the pseudo-code for the first compression algorithm.

Figure 8.1 shows the F0 values and cubic regressions of those values for the pitch contours of the following utterance and summary pair:

Original: *So given these um these features or or these these examples um critical examples which they call support f- support vectors then um given a new example if the new example falls um away from the boundary in one direction then it's classified as being a part of this particular class*

Compression: *So given these features or these examples critical examples which they call support vectors then given a new example if new example falls boundary in one direction then being a part of this particular class*

It is worth noting that these cubic regressions are highly stylized versions of the pitch contours, and that in reality the pitch data is much noisier than the regressions indicate. Furthermore, there are a variety of factors that could cause a given compression to have a substantially different pitch contour than the overall dialogue act contour, e.g. skipping disfluent fragments and removing filled pauses.

The second method is more crude and does not depend on recognizing phrase boundaries. Instead, the pitch contour for the entire dialogue act is represented as a vector of F0 values. Compression proceeds by deleting words one at a time, based on how large an effect each word's deletion has on the pitch contour. For each iteration of the procedure, each word has its F0 values deleted from the pitch vector and replaced with interpolated values between its former neighbouring words. This new

```
for each: dialogue act
    create vector of dialogue act F0 values
    for each: word
        delete word's F0 values from dialogue act vector
        vector_word = interpolate missing F0 values
        calculate cosine(vector_word , original vector)
    current length = full length
    while: current length > desired length
        remove arg max_word f(word) = cosine(vector_word , original vector)
        current length -= 1
```

Table 8.2: Second Compression Algorithm

pitch vector is then compared with the original pitch vector by using cosine similarity. The word with the highest cosine similarity is deleted, as the removal of its F0 values had little effect on the overall pitch contour. Again, the procedure continues until the desired length is reached. Table 8.2 gives the pseudo-code for the second compression algorithm.

Essentially, the two prosodic methods are working from opposite directions, one iteratively selecting words while the other is iteratively eliminating words. There are significant procedural differences, however, as the latter method does not use phrasal information and thus would not ignore short fragments as the former method would. This second method also relies on overall pitch vector similarity, which may not be as reliable as measuring slope at the phrasal and word levels.

Figure 8.2 shows cubic regressions for the pitch contours of the following utterance and summary pair:

Original: *And the interesting thing is that even though yes it's a digits task and that's a relatively small number of words and there's a bunch of digits that you train on it's just not as good as having a a l- very large amount of data and training up a a a nice good big HMM*

Compression: *And interesting thing is that though yes it's digits task and that's relatively small words and there's bunch digits you train on it's just not good as having a large amount and training up a nice good big HMM*

8.1.3.2 Simple Text Method

For the second evaluation scheme described below, we implement a simple text compression method for comparison. As in the methods described above, we began by deleting filled pauses and repetitions. We then assign each word in the dialogue act

a *tf.idf* score, a metric which gives high ranks to words that are frequent within a document but rare across multiple documents. We select the words with the highest *tf.idf* scores until the desired compression length is reached. This text compression method is quite simple but nevertheless one would have a reasonable expectation of high informativeness using this method.

8.1.3.3 Baseline

To assess baseline performance, we randomly select the desired number of words, using the same compression level, and present them in the original order.

8.1.3.4 Gold Standard

The gold standard for compression is human-authored compressions. Manual compressions are made with a compression rate between 60% and 70%. The manual compressions are restricted to using only words from the original dialogue act and are presented in the original order, as with the automatic methods. The slightly wider window for the compression rate is because it is not feasible to require human annotators to compress an utterance to a precise percentage of the original. The manual compressions for these experiments were created by a single human annotator.

8.1.4 Evaluation

Two methods of evaluation are carried out, the first being a subjective analysis using human annotators who rate each compression on two criteria, and the second being a measure of edit-distance to a gold-standard compression. The text compression method was not implemented until after the human evaluation was complete, and so it is only included in the edit-distance evaluation.

Thirty dialogue acts from the ICSI corpus are chosen which were output from the summarizer described in (Murray et al., 2006), which represents early work on the ICSI corpus using a much smaller corpus of lexical and prosodic features than the database described in Chapters 5 and 7. These dialogue acts average about 27 words in length. The content of the dialogue acts is quite technical, and though it would have been possible to select less technical and shorter dialogue acts, we are fundamentally concerned with how our compression method performs on actual summarizer output.

8.1.4.1 Subjective Evaluation

Five human judges are presented with the output of four compression methods on the test set, for a total of 120 compressions to be evaluated. These four methods are random baseline compressions, human-authored gold-standard compressions, and the two prosodic compression methods. The judges are asked to rate each compression for two criteria, informativeness and readability. The ratings are made on a 1-5 Likert scale with 1 being 'Very Poor' and 5 being 'Very Good.'

8.1.4.2 Informativeness

When rating a given compression in terms of its informativeness, judges are asked to keep in mind whether the compression retains the most important parts of the original utterance and refrains from including irrelevant or unnecessary parts of the original. They are instructed that this is a distinct and separate rating from readability, so that a compression may score high on informativeness and still do very poorly on readability.

8.1.4.3 Readability

When rating a given compression in terms of its readability, judges are asked to consider whether the compression seemed grammatical and fluent relative to the original and whether the compression is generally readable. The term *relative* is included in the instructions because a compression which is an ungrammatical fragment should not be scored very low if the original utterance was also an ungrammatical fragment, for example.

8.1.4.4 Edit Distance

The second method of evaluation is edit distance, which utilizes our human-authored compressions as a gold-standard for an objective comparison. The edit distance between two strings is defined as $1-(I+D+S)/R$, where R is the number of words in the reference string and I, D and S are insertions, deletions and substitutions, respectively. This metric thus objectively measures how close an automatically compressed string comes to the ideally compressed string. For this evaluation, four compression approaches are measured against the reference string, with the four approaches being random, text-based, and two prosodic approaches.

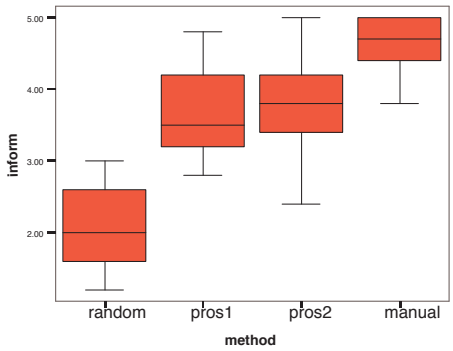


Figure 8.3: *Informativeness Scores for Four Compression Methods*

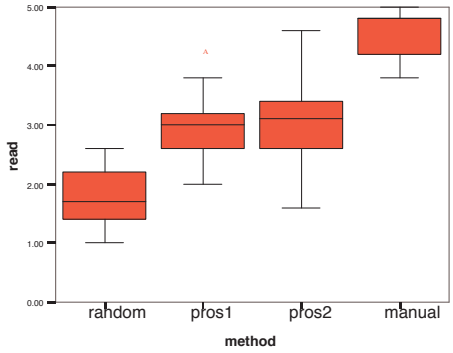


Figure 8.4: *Readability Scores for Four Compression Methods*

8.1.5 Results

8.1.5.1 Subjective

Figure 8.3 shows the averaged informativeness scores for the four compression methods. The inter-annotator agreement is very good, with the correlation of macro-averaged scores above 0.9 for each annotator pair. The manual compressions are rated significantly higher than the others ($p<0.05$), with an average informativeness score of 4.65. Both of the prosodic approaches are significantly better than random ($p<0.05$) but are not significantly different from one another. The first prosodic approach has an average informativeness score of 3.69 and the second prosodic approach has an average of 3.82. The random compressions average 2.08 in terms of informativeness.

Figure 8.4 shows the averaged readability scores for the four compression methods. The inter-annotator agreement is again very good, with correlations above 0.9 for each

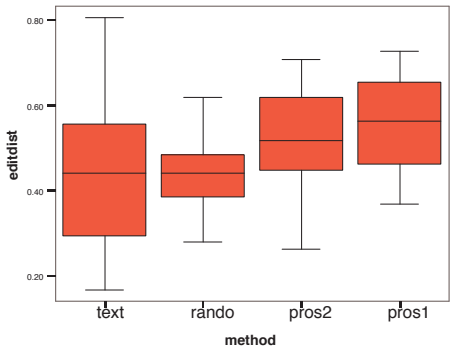


Figure 8.5: *Edit Distance for Four Compression Methods*

annotator pair. The significant effects are the same as those of the informativeness scores, with the manual compressions rating significantly higher than the other approaches ($p<0.05$) and the prosodic approaches being significantly better than random ($p<0.05$) but not significantly different from one another. The manual compressions have an average readability score of 4.6, the first prosodic approach averages 2.93, the second prosodic approach averages 3.15, and the random compressions average 1.77 in terms of readability. While the random and prosodic approaches have readability scores significantly lower than their informativeness scores, the manual compressions score comparably on both readability and informativeness. It's clear that human judges are able to separate the two criteria when giving their ratings.

8.1.5.2 Edit Distance

Figure 8.5 shows the results of the edit distance metric, in which the manual gold-standard compressions are compared with the random and prosodic approaches, as well as a simple *tf.idf* approach. The most striking aspect of these results is that the *tf.idf* method performs only at the level of the random method. The prosodic approaches are significantly better ($p<0.05$), with an average edit distance of 0.56 and 0.53, respectively. The *tf.idf* and random approaches each have an average edit distance of 0.44. It can be noted that there is a large amount of variance with the *tf.idf* approach, sometimes performing very well and other times failing completely.

8.1.6 Conclusion

This section has presented a novel method of compressing utterances by preserving the pitch contour of the original within the compressed version. This compression method is meant to be robust to the disfluencies and ungrammaticalities of meeting speech, and the results are encouraging. We report the findings of a pilot study evaluating two implementations of this approach. Based on both subjective and objective evaluation metrics, the prosodic approaches are far better than random compression. Objective evaluation using edit-distance also shows the prosodic methods outperforming a keyword-based compression approach. Relative to human-authored gold-standards, the readability of the prosodic compressions suffers but there are quite high levels of informativeness.

Though the second prosodic method was thought to be cruder than the first, it performs slightly but not significantly better in terms of both readability and informativeness. Future work may combine the two methods in order to optimize the compression results.

This is very early work on compression for meeting dialogue acts, but it does indicate that there is a role for prosody in such a task. Future work would likely benefit from the inclusion of additional features such as ASR confidence scores and n-gram language modelling to increase informativeness and readability.

8.2 Towards Online Speech Summarization

8.2.1 Introduction

The majority of speech summarization research has focused on extracting the most informative dialogue acts from recorded, archived data. However, a potential use case for speech summarization in the meetings domain is to facilitate a meeting in progress by providing the participants - whether they are attending in-person or remotely - with an indication of the most important parts of the discussion so far. This requires being able to determine whether a dialogue act is extract-worthy before the global meeting context is available. This section introduces a novel method for weighting dialogue acts using only very limited local context, and shows that high summary precision is possible even when information about the meeting as a whole is lacking. The novel online summarization method is shown to significantly increase weighted f-scores compared with a method using no contextual information.

When applying speech summarization to the meetings domain, the goal of most research has been to extract and concatenate the most informative dialogue acts from an archived meeting in order to create a concise and informative summary of what transpired. Such summaries are analogous to the traditional manual minutes of a meeting, and are relevant to use cases such as a person wanting an overview of a meeting they missed, or a person wanting to review a meeting they attended, as a mental refresher. However, there are many use cases that go beyond the scenario of a user accessing an archived meeting. For example, someone might join a meeting halfway through and require a method of catching up on the discussion without disturbing the other participants. A second example is a person who is remotely monitoring a meeting while attending to another task, with the intention of joining the group discussion when a certain topic is broached. These use cases require the development of online summarization methods that classify dialogue acts based on a much more limited amount of data than previously relied upon.

This section introduces effective methods for scoring and extracting dialogue acts based on examining each candidate's immediate context. A method of *score-trading* is introduced and described wherein redundancy is reduced while informativeness is maximized, thereby significantly increasing weighted f-scores in our evaluation.

8.2.2 Weighting Dialogue Acts

This section describes three methods of scoring and extracting dialogue acts, the first of which relies on a simple term-score threshold, and the second two of which rely on a more complex score-trading system within the dialogue act's immediate context.

8.2.2.1 Residual IDF

The experiments described in Chapter 4 have shown *ridf* to be superior to IDF on this data. Our first method of extraction then is to simply sum *ridf* term-scores over each dialogue act and extract a given dialogue act if it exceeds a pre-determined threshold. Based on using various thresholds on a separate development set of meetings, a threshold of 3.0 is used for the experiments below. *ridf* scores were calculated using a collection of documents from the AMI, ICSI, MICASE and Broadcast News corpora, totalling 200 speech documents (AMI test set meetings were excluded). We cannot use *su.idf* or *tf.idf* for this set of experiments, as they require a great deal of meeting context, if not the entire meeting, in order to calculate the term-weights. In contrast,

ridf relies solely on frequencies from the document collection without reference to term-frequency in the document at hand.

8.2.2.2 Score-Trading

The previously described method uses no knowledge of dialogue act context, and therefore does not address redundancy or importance relative to neighboring dialogue acts. A dialogue act is simply extracted if it scores above a given threshold. In contrast, the following two methods use a limited amount of context in order to maximize informativeness in a given region and to reduce redundancy, via a simple score-trading scheme.

For each dialogue act, we examine the ten preceding and ten subsequent dialogue acts. For each unique word type in that 21-dialogue-act window, we total its overall score (its *ridf* score times its number of occurrences in that window) and reapportion that overall score according to the relative informativeness of the dialogue acts containing the term. For example, if the word 'scroll' has an *ridf* score of 1.2 and it occurs twice in that window, in two different dialogue acts, it has a total score of 2.4. If one of the dialogue acts containing the term 'scroll' has a dialogue act score of 5.0 and the other has a dialogue act score of 3.0, the overall term score is apportioned in favor of the former dialogue act, so that is receives a revised term score of 1.5 and the latter receives a revised term score of 0.9. As a result, the dialogue act score for the former has increased while it has decreased for the latter. This method of score-trading places the burden of carrying that term's information content onto the more generally informative dialogue acts, which also has the effect of reducing redundancy. Figure 8.6 illustrates the basic premise behind this scheme.

A dialogue act's initial score, or Ascore, is simply the sum of its constituent words' *ridf* scores:

$$Ascore(d) = \sum_{i=1}^{W} ridf(t_i)$$

where W is the number of words in the dialogue act. The revised term-score for word t in dialogue act d is given by

$$Trade(t,d) = ridf(t) \cdot N(t) \cdot \left(\frac{Ascore(d)}{\sum_{i=1}^{M} Ascore(i)}\right)$$

where $ridf(t)$ is the original *ridf* score for the term, $N(t)$ is the number of times that the term t appears in the context window, and M is the number of dialogue acts in the window that are indexed by term t.

Figure 8.6: *Score-Trading Between Dialogue Acts*

A dialogue act's Bscore is then the sum of its revised term-scores:

$$Bscore(d) = \sum_{i=1}^{W} Trade(t_i, d)$$

After deriving the Bscore score, the dialogue act in question is extracted if it satisfies the case

$$Bscore >= 3.0$$

The second score-trading method is similar to the first, but a dialogue act is extracted if it satisfies the formula

$$(2 \cdot Bscore) - Ascore >= 3.0$$

where Ascore is the original score and Bscore is the adjusted score. The reasons motivating this latter method are twofold. First, a dialogue act's adjusted score (i.e. Bscore) may still be below the 3.0 threshold, but if it has increased significantly compared to the Ascore, that indicates its importance in the local context and we want to increase its chances of being extracted. Second, a dialogue act's adjusted score may be above 3.0 but it is well below its original Ascore, indicating that it has lost informativeness and may well be redundant in the local context. As a result, we want to reduce its chance of being extracted.

8.2.3 Experimental Setup

For this set of experiments we use the AMI meeting corpus test set, comprised of 20 meetings total. For our evaluation, we rely on weighted precision/recall/f-score as used in previous chapters and described in detail in Chapter 3 Section 3.5.4.1 (page 33).

The generated summaries range between 600 and 3000 words in length, as the meetings themselves greatly vary in length. Unlike summarization of archived meetings, here we do not specify a set summary length in advance since the length of the meeting is not known beforehand. It would be possible to set an extraction ratio and/or to have the resultant summaries revised to fit a particular length requirement once the meeting has finished, but here we simply decide whether or not to extract each dialogue act candidate without consideration of the summary length at that point.

sys	man-prec	man-rec	man-fsc	asr-prec	asr-rec	asr-fsc
ridf	0.608	0.286	0.382	0.612	0.276	0.374
trade	0.611	0.295	0.391	0.610	0.285	0.383
tdiff	0.603	**0.305**	0.399	0.605	**0.295**	**0.392**

Table 8.3: Weighted Precision, Recall and F-Scores
ridf=DA extracted if Ascore >= 3.0, **trade**=DA extracted if Bscore >= 3.0, **tdiff**=DA extracted if Bscore - (Ascore-Bscore) >= 3.0

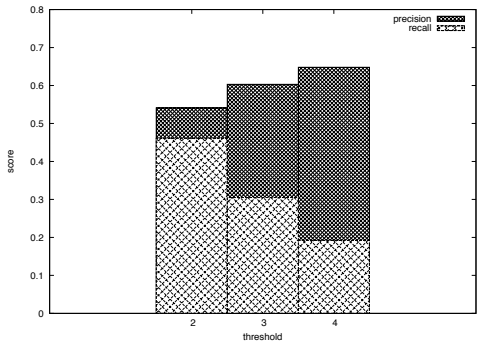


Figure 8.7: *Score-Trading at Multiple Thresholds*

8.2.3.1 Results

Table 8.3 presents the weighted precision, recall and f-scores for the three approaches described above. One of the most surprising results is that the weighted precision in general is not drastically lower than the scores found when creating very brief summaries of archived meetings. For example, in Chapter 4, creating 700-word summaries of the same test set using *ridf* yielded an average weighted precision of 0.66 (page 48). All three online approaches presented here have average weighted precision around 0.61. This is particularly surprising and encouraging given that these summaries are on average much longer than 700 words.

The third approach, labelled **tdiff** in Table 8.3, is superior in terms of f-score on both manual and ASR transcripts. *ridf* performs the worst on both sets of transcripts, and the second approach labelled **trade** is in-between. Significant results in the table are presented in boldface. The method **tdiff** achieves significantly higher recall than the other two methods on manual transcripts, and both recall and f-score are significantly higher on ASR (paired t-test, $p<0.05$). The most encouraging result of this third approach is that it is able to significantly increase recall without significantly reducing

sys	man-prec	man-rec	man-fsc	asr-prec	asr-rec	asr-fsc
trade	0.599	0.291	0.386	0.608	0.291	0.388
tdiff	0.589	**0.306**	**0.398**	0.593	**0.304**	**0.398**

Table 8.4: Weighted Precision, Recall and F-Scores (Offline)
trade=DA extracted if Bscore >= 3.0, **tdiff**=DA extracted if Bscore - (Ascore-Bscore) >= 3.0

precision.

Having determined the effectiveness of the third approach, we subsequently run this score-trading method at multiple thresholds of 2.0, 3.0 and 4.0 to gauge the effect on weighted precision, recall and f-score. The results are displayed in Figure 8.7. A threshold between 2 and 3 results in a good balance between recall and precision, while a threshold of 4 results in drastically lower recall and only slightly higher precision.

The score-trading results reported so far stem from an implementation of the method that has an algorithmic delay of 10 dialogue acts. We are interested in what benefit, if any, could be gained by increasing the algorithmic delay and thereby increasing the amount of context used. The two score-trading approaches are therefore run fully offline, so that the context for each dialogue act is the entire meeting (the first approach, based simply on *ridf* results, is the same online versus offline since it does not use context). Because there is a larger amount of score-trading when using all meeting dialogue acts for comparison, a given dialogue act would have to be very informative in order to have its overall Ascore increase. The expectation is that running this method offline would result in higher precision and perhaps lower recall. Table 8.4 presents the weighted precision, recall and f-scores for the offline systems. The third approach, labelled **tdiff** in Table 8.4, is again superior to the second approach, labelled **trade**, with significant differences between the two in terms of recall and f-score on both manual and ASR transcripts. However, neither approach is significantly different when run offline versus online. The trend is for precision to be slightly lower when run offline and recall to be slightly higher, the opposite of what was expected.

8.2.4 Discussion

The results above show that the score-trading scheme is able to significantly increase recall and f-score with no significant decrease in precision. More specifically, it allows

us to reject dialogue acts that may have scored highly but were redundant compared with similar and more informative neighboring dialogue acts, and allows us to retrieve dialogue acts that may have scored below the threshold originally but subsequently had their scores adjusted based on local context.

In general, it is interesting that high precision is attained via these three methods that use either no context or only local context. As mentioned earlier, previous experiments on creating very concise summaries using global information about the meeting achieved weighted precision of only a few points higher than these results. It turns out that restrictions such as the inability to create an overall ranking of dialogue acts in a meeting and inability to rely on term-frequency information are not severely detrimental to the ultimate results. In Chapter 4 (e.g. pages 48 and 51), we found that term-weighting approaches that do not rely on overall term-frequency in the meeting can perform very competitively, and this experiment yields a similar conclusion.

A related finding is that there is no benefit to running the score-trading methods completely offline, using the entirety of the meeting's dialogue acts as context. In fact, precision results were slightly better when examining only the limited context. It may be that dialogue acts sharing some of the same terms and existing within proximity to each other tend to be more similar than dialogue acts sharing some of the same terms but existing at various locations spread throughout the meeting. In that case, score-trading between ostensibly similar dialogue acts would not always be beneficial if the examined context is too large. This relates to work by Galley (2006), who proposed a restricted form of Pyramids evaluation based on the observation that words that are similar but occur in different parts of the meeting can have very divergent meanings.

While the score-trading methods outperform the simple *ridf* threshold method, with the third summarization system performing the best, it would seem that the methods are complementary. Because the *ridf* method requires no contextual information, a dialogue act can be immediately extracted or rejected on a preliminary basis. Once the subsequent context for a dialogue act becomes available, that decision can be revised based on score-trading. User feedback could provide a further source of input for such dynamic summary creation.

Score-trading is similar in spirit to MMR (Carbonell & Goldstein, 1998), described in detail in Chapter 5 Section 5.2 (page 64), in that they both work to heighten informativeness and reduce redundancy in the summary. Whereas MMR penalizes a given sentence with a redundancy score based on similarity to already-extracted sentences, we compare each candidate dialogue act to its surrounding dialogue acts, and a dia-

logue act can have its score decrease, increase or remain the same based on how generally informative it is and whether the surrounding dialogue acts have term overlap with the candidate.

8.2.5 Conclusion

This section has introduced a novel method for the online summarization of spoken dialogues, using a score-trading scheme intended to reduce redundancy and to develop a more subtle view of informativeness. By looking at informativeness beyond the level of the dialogue act and examining local context around the candidate dialogue act, we are able to locate words that are generally informative in a local region of the meeting transcript and to place the burden of carrying those words' informativeness onto the most informative dialogue acts in that region. An encouraging finding for the prospect of online meeting analysis is that weighted precision scores are not drastically lower than the precision scores found in previous work on very concise summarization of archived meetings, even when the recall of the summaries contained herein is much higher. Running the score-trading methods offline does not result in any added benefit compared with using only a small amount of context and executing the method online.

8.3 Summarization Without Dialogue Acts

8.3.1 Introduction

In previous chapters, our summarization systems have relied on dialogue acts as input, using those segments as the units of extraction. In this section, we briefly consider the use of *spurts* rather than dialogue acts as our summary units (Shriberg et al., 2001). A spurt can simply be defined as a region where a meeting participant is speaking continuously, with boundaries determined by pause information. A primary benefit of using spurts rather than dialogue acts is that we can quickly segment the speech stream into meaningful units without time-consuming dialogue act segmentation. This is of particular importance for online summarization as described in the previous section. Spurt segmentation may also result in units of finer granularity than dialogue acts and allow us to more accurately pinpoint informative regions of the meeting.

8.3.2 Spurt Segmentation

In defining spurts, we rely entirely on pauses and filled pauses for determining the unit boundaries. This is in contrast to most work on dialogue act segmentation, where prosodic features along with n-gram language models are used for segmentation (Ang et al., 2005; Dielmann & Renals, 2007). Taking automatically speaker-segmented ASR output as our input, we place a spurt boundary at any location where the inter-word pause for a speaker is 400 ms or longer, or where there is a pause of at least 200 ms plus a filled pause such as "um," "uh," or "erm." Once we have segmented the speech stream of each speaker in the meeting, the final input to the summarization system is the list of spurts ordered so that they are monotonically increasing according to start-time.

8.3.3 Experimental Overview

These spurt-based experiments are performed on the AMI corpus test set.

Once we have the input format described above, summarization proceeds simply by scoring each spurt using the *su.idf* metric, identical to the process described in Chapter 4 Section 4.1.3 (page 46). Each spurt's score is calculated as the sum of its constituent word scores. We then rank the spurts according to their scores and extract until we reach the length limit of 700 words.

Throughout most of this thesis, we have relied on weighted precision/recall/f-score for our evaluation metrics, using multiple human extractive annotations of dialogue acts. Now that the summarizer no longer uses dialogue acts as its summary units, we have to rely on other evaluation metrics. For this purpose, we use the ROUGE-2 and ROUGE-SU4 n-gram metrics (Lin, 2004), which calculate bigram and skip bigram overlap between automatic and multiple reference summaries.

For comparison, we include human summaries of the same length, 700 words, choosing one annotator at random for each meeting and extracting their most-linked dialogue acts until reaching the length limit. These human summaries are then also compared with human gold-standard abstracts using ROUGE.

8.3.4 Results

Table 8.5 lists the ROUGE-2 scores for the AMI test set meeting summaries, for both the automatic spurt-based approach described above and human-level performance.

Meet	ASR-Spurts	Human
ES2004a	0.02657	0.05105
ES2004b	0.01770	0.01735
ES2004c	0.03994	0.02675
ES2004d	0.01102	0.01362
ES2014a	0.06946	0.08037
ES2014b	0.03252	0.03518
ES2014c	0.06032	0.07938
ES2014d	0.05168	0.06133
IS1009a	0.10370	0.14720
IS1009b	0.02184	0.06278
IS1009c	0.03873	0.10256
IS1009d	0.06166	0.08995
TS3003a	0.04813	0.04558
TS3003b	0.07564	0.04234
TS3003c	0.06742	0.06541
TS3003d	0.04843	0.05155
TS3007a	0.08180	0.08254
TS3007b	0.01933	0.01591
TS3007c	0.04792	0.06069
TS3007d	0.02420	0.02417
AVERAGE	**0.047**	**0.058**

Table 8.5: ROUGE-2 Scores for Spurt Summarization and Human Summarization

We find that according to ROUGE-2, not only does performance not decrease when using simple spurt segmentation instead of dialogue act segmentation, the scores are actually higher than the ROUGE-2 scores reported in Chapter 7 Section 7.4.3.1 (page 140). The ROUGE-2 average for the meta summaries applied to ASR described in that chapter is 0.041 compared with 0.047 here, a significant result according to paired t-test ($p<0.05$).

Table 8.6 lists the ROUGE-SU4 scores for the spurt-based summaries and the human summaries. The average for the spurt-based approach is 0.079, which again is significantly better than the highest ROUGE-SU4 scores reported in Chapter 7 (page 140), 0.070 ($p<0.05$). We also find that the average for the spurt-based method approaches human-level performance on this metric. On many meetings it is in fact superior to human performance.

The following excerpt of the summary for meeting TS3003c shows that many of the extracted spurts are quite short, often less than 10 words, compared with the higher word counts for the AMI and ICSI summary dialogue acts in Chapter 5 Section 5.6 (page 77):

Speaker D: Uh the remote control and the docking station

Meet	ASR-Spurts	Human
ES2004a	0.04255	0.06016
ES2004b	0.04845	0.05664
ES2004c	0.06145	0.07204
ES2004d	0.04263	0.04812
ES2014a	0.09303	0.10463
ES2014b	0.07475	0.07813
ES2014c	0.09769	0.10030
ES2014d	0.08080	0.07982
IS1009a	0.15810	0.15256
IS1009b	0.06884	0.08556
IS1009c	0.06422	0.11776
IS1009d	0.10325	0.12989
TS3003a	0.06924	0.05704
TS3003b	0.12962	0.07534
TS3003c	0.10156	0.10064
TS3003d	0.06387	0.07626
TS3007a	0.09938	0.09572
TS3007b	0.05804	0.05687
TS3007c	0.08521	0.07990
TS3007d	0.05244	0.05595
AVERAGE	**0.079**	**0.084**

Table 8.6: ROUGE-SU4 Scores for Spurt Summarization and Human Summarization

Speaker D: Docking station and small screen would be or main points of interest because this would be

Speaker D: Advise that it should be remote control on the docking station should be telephone

Speaker D: So you could imagine that uh the remote control be standing up

Speaker D: Design where the remote control just lies

Speaker D: And grey black colour for the

8.3.5 Discussion

The reason that the spurt-based approach performs better than the dialogue-act based approach according to ROUGE seems to be that there is a finer level of granularity. For the AMI test set, there are on average nine fewer dialogue acts extracted for each meeting compared to spurts extracted. The spurts simply tend to be shorter, and so we can extract more of them. The mean word count of a dialogue act in the AMI test set meetings is 6.3 (s.d. 7.1), compared with 5.4 (s.d. 6.6) for spurts (interestingly, however, the average of the *longest* spurt from each meeting is very slightly higher than

the average of the longest dialogue act). Since our units are a finer granularity it may be that we can more easily separate the informative and non-informative portions of the transcript. For example, with dialogue acts we might extract a very long dialogue act because it has several high-scoring words, but in fact there may be only one part of the dialogue act that is particularly relevant and the remainder would simply be included because it is one extraction unit.

Of course, one solution to this problem is to compress dialogue acts after extraction, and the first section of this chapter described one set of compression experiments. However, a certain amount of compression would be unnecessary if we began with a finer granularity for our extraction units. It is somewhat of a roundabout process to segment dialogue acts, extract the most informative ones, which tend to be longer units, and then compress them, compared with simply using finer extraction units to begin with. Compression is still very useful, especially when the informative portions of the extraction unit are spread throughout the unit with intervening uninformative words or phrases, but using spurts may decrease our need to carry out any further compression.

8.4 Conclusion

This section has described three areas of further research on extractive summarization and discussed preliminary methods of addressing the inherent challenges. For automatic compression, we described prosody-based methods that outperform a simple text-based approach for compressing a set of dialogue acts from the ICSI corpus. Compression is an active area of research in and of itself and the methods presented here are preliminary and fairly simple, but they illustrate to some extent the usefulness of prosodic features for this task. Other possible features of interest for dialogue act compression are n-gram language model probabilities and ASR confidence scores, and a competitive system could perhaps combine such features with prosodic information for robust compression results.

We then addressed the challenge of online summarization, wherein we must decide whether or not to extract a dialogue act before we have the full context of the meeting. We introduced a score-trading mechanism by which we adjust dialogue act scores based on the immediate surrounding context in the meeting. We showed that even without access to the full meeting, we are able to extract dialogue acts with a high level of precision. This finding is particularly interesting within the scope of the AMIDA project, where meetings are automatically analyzed in as close to real-time as

possible. Two variations on score-trading were both found to be effective, increasing the weighted f-score compared with extraction based on summed *ridf* scores alone.

In the third section of this chapter, we discussed spurt-based segmentation, carrying out summarization with extraction units segmented by pause and filled pause information. We find that by using spurts, our ROUGE scores surpass the ROUGE scores from using dialogue act segments. Additionally, we find that the ROUGE results for spurt-based summarization approach human performance on the AMI corpus meetings. The findings of the spurt results have ramifications for the results in the prior two sections of this chapter. First of all, we hypothesize that using a finer level of granularity in our extraction units will make us less reliant on needing to carrying out further compression. And second, spurt-based segmentation will be very useful for online summarization, wherein the speech stream must be segmented as quickly as possible and full dialogue act segmentation is not feasible.

Finally, we can briefly mention other challenges and directions for speech summarization in the coming years. In general, it is expected that the speech summarization community in the immediate future will focus increasingly on moving beyond simple extractive summarization, as will the summarization community in general. In Chapter 7 we have laid groundwork that will hopefully inform and aid future research in making extractive summaries more abstractive, or creating hybrid summaries.

The unique nature of group multi-modal interaction means that multi-modal *summaries* can be generated to convey the information contained in a meeting. Future summarization work in this domain may increasingly look at combining edited video, audio, and transcripts with selected slides, screenshots and notes in order to create complex summaries from multiple sources. These multi-modal information sources can be exploited not only in the meeting browser, but as features in the summarization system itself. For example, note-taking behaviour by the meeting participants might be particularly indicative of the presence of relevant information at that point in the discussion.

Work on the summarization of meetings will also look increasingly at complex interactions involving remote participants who may be linked via video or telephone. As remote conferencing equipment becomes more commonplace, it will become increasingly rare to find that all meeting participants will be attending the discussion in person. Technologies that facilitate these long-distance conversations both during and between meetings will be vital.

More generally, in the coming years speech summarization research will likely

expand into new speech domains. Most work to date has been carried out on broadcast news, lectures, telephone speech or meetings. In the near future the popularity of podcasts and audio-based discussion forums may merit research in those areas.

It has been widely agreed upon that summarization researchers should incorporate extrinsic evaluations as much as possible rather than solely relying on intrinsic measures, and the coming years may see the development of widely accepted and adopted extrinsic schemes for evaluating summaries of this type of data. In Chapter 6 we have described one such extrinsic evaluation, the *decision audit* task, which hopefully can inform the eventual adoption of a standardized extrinsic task in the community.

The speech summarization community would greatly benefit from an annual meeting along the lines of the Document Understanding Conference, which, for the text summarization community, has allowed researchers to establish the state of the art and compare numerous systems on benchmark tasks. Such a conference would likely expedite development in this field and introduce researchers to the current relevant work going on at institutions other than their own. Perhaps no other single step would increase the quality of speech summarization in the immediate future as much as this would.

Chapter 9

Conclusion

9.1 Discussion

In this thesis we have examined the task of extractive summarization of spontaneous multi-party speech, and specifically researched the usefulness of various multi-modal features for several stages of the summarizaton process. The experimental hypothesis has been that utilizing the multitude of features available for this type of data would be beneficial compared to simply treating the summarization task as a text summarization task with a noisy transcript. We have repeatedly demonstrated this hypothesis to be true, showing that a rich variety of lexical, prosodic, structural, and speaker features yield optimal results for the overall task.

The most important contributions of this research are four-fold, which we summarize in order of the thesis structure. First, we have compared several term-weighting approaches in terms of summarization performance on meeting data. Our first novel term-weighting method, *su.idf*, which relies on differing term usage among speakers, performs at state-of-the-art levels compared with the techniques imported from text IR. A second novel method, *twssd*, relies on speaker and structural correlates of terms and performs competitively without any recourse to term-frequency or collection-frequency information. We also found that these novel metrics differ between corpora, with the former performing better on the AMI data and the latter rating better on the ICSI data. In Chapter 5, we used *su.idf* and *tf.idf* as the vector weights in a variety of summarization systems and found *su.idf* to be consistently superior. This portion of the research will inform future work on speech summarization, as choosing a term-weighting metric is a vital first development step for most summarization systems and there are numerous metrics from which to choose.

Second, we have conclusively determined the most effective sets of features for summarizing meeting data, using multiple corpora, so that future research on this data will be well-informed in regards to deciding which individual features and general feature sets merit inclusion in the research. The features analysis is the most comprehensive that we have seen for these two meeting corpora in regards to automatic summarization research. More importantly, we have illustrated the importance of investigating a wide variety of features in order to achieve optimal results. In Chapter 5, we described several unsupervised summarization techniques applied to this data, as well as supervised methods utilizing a variety of multi-modal features for this data. We found that we could achieve the best extraction results according to weighted f-score by using a combination of lexical, prosodic, speaker, structural and length features. We also found that using certain feature subsets alone can yield good summarization results. Using only prosodic features, for example, allows us to create decent quality summaries, allowing for the possibility of creating audio-to-audio summaries without automatic speech transcription. It was also repeatedly found that summarization results did not deteriorate on the ASR-aligned databases. The summarization techniques in general are very robust to moderately high WERs for these corpora. In the same chapter, we characterized how informative and uninformative dialogue acts are realized in terms of their feature correlates.

Third, we have presented a large-scale extrinsic evaluation for summarization in this domain. While we relied on intrinsic measures of summarization quality in Chapters 4, 5 and 7, in Chapter 6 we described an extrinsic evaluation for comparing summarization types. The specific formulation of the task was a *decision audit*, where users had to review several archived meetings in order to satisfy a complex information need, utilizing different information sources in each condition. We found that the extractive summaries were highly rated in terms of user satisfaction, human objective and subjective evaluations, and in terms of efficient user browsing and writing behaviour during the task. Though users rated summaries of ASR transcripts substantially lower in terms of satisfaction, they were able to adapt and modify their browsing behaviours by using the summary dialogue acts as indices into the meeting and then relying much more on audio/video disambiguation. Users find these summaries to be intuitive and efficient for browsing meetings in time-constrained situations. It is widely agreed upon in the summarization community that such extrinsic measures should increasingly supplement the intrinsic measures that are used for development purposes. The impact of the *decision audit* evaluation is significant in that it justifies research on

extractive speech summarization by showing that such summary types are useful for navigating through the meeting data, and in that it provides a possible model for future large-scale summarization evaluation within the research community. Specifically, if speech summarization researchers organize a regular conference, such an extrinsic task could form one of the system evaluation components.

Four, we have presented work that begins to move the speech summarization state-of-the-art further down the extractive-abstractive continuum. Though the research described in Chapter 7 was still firmly in the extractive paradigm, with the aim of extractive informative dialogue acts as our summary units, the purpose of those experiments was to find high-level meta dialogue acts in the meeting: areas where the speakers are referring to the meeting itself, often in terms of decisions, goals or problems that were encountered. We found that we could discern such dialogue acts from other dialogue acts by using a diverse multi-modal set of features, including abstractive cuewords. These dialogue acts are realized distinctly from the dialogue acts labelled as "extractive" in Chapter 5, in terms of their prosodic and structural correlates. We evaluated these new "meta" summaries using three metrics – weighted precision using the new extractive labels, weighted precision using the old extractive labels, and ROUGE – and found the new summaries to be superior on each measure. By creating summaries that include as much high-level perspective as possible while relating the informative portions of the meeting, we end up with summaries that are more abstractive in quality than previous extractive summaries. This aspect of the work will hopefully inform future research on abstractive summarization and hybrid summarization.

At multiple points in this research, we found our summarization systems performing near or at human-level performance on the given tasks. In Chapter 5, weighted precision scores are level between human and automatic summarizers on the ICSI corpus. For the AMI corpus, though the average precision scores are lower for the automatic summarizers overall, on numerous individual meetings the automatic systems meet or exceed human performance according to this metric. In Chapters 7 and 8, which incorporate ROUGE metrics, we find automatic performance nearing human-level performance according to ROUGE-2 and ROUGE-SU4 on the AMI corpus test set. However, the system is strongest in terms of precision and weaker in terms of recall, as it tends to extract longer dialogue acts for inclusion in our brief summaries. Future work will aim to counteract the system's dependence on length features.

A limitation of most of the summarization systems described herein is that they focus on classifying a candidate sentence according to features of that candidate sen-

tence, such as prosodic cues and term-weights, with less regard for features of the sentence context. For example, further investigation could determine that informative sentences tend to be preceded or succeeded by certain patterns in the signal or the text. The score-trading method described in Chapter 8 Section 8.2 (page 158) is one attempt at examining the candidate sentence's context. Predictive features in particular could be beneficial for the use case where somebody remotely monitoring the meeting wants to be alerted when a subject is about to be broached or a topic is about to shift.

This research for the most part has not addressed the treatment of disfluencies in spontaneous speech, other than the removal of filled pauses. Further work in this area could greatly increase the coherence and readability of extractive summaries. In a similar vein, the derivation and use of confidence scores for the recognition output on these meeting corpora would likely yield both greater informativeness and readability.

9.2 Conclusion

Meetings are increasingly a ubiquitous part of people's lives, and technologies as described above and implemented in a browser framework will allow individuals to make more efficient use of their time between meetings and during meetings, whether they are attending in person or remotely. The discussions that happen in such meetings are unique in that they often exhibit low information density, multi-modal information sources, and distinct speaker roles and structural characteristics, which together warrant the application of extractive technologies that incorporate these features. We have shown that such summarizers can efficiently discern the most informative portions of the meeting from the remainder, and that said systems are capable of challenging human-level precision on this task.

Appendix A

Decision Audit Documents

This appendix provides materials used in the decision audit experiments, including the pre-questionnaire given to the participants, written instructions for the task, and the post-questionnaire with Likert-scale statements.

A.1 Pre-Questionnaire

Pre-Questionnaire

Please answer the following questions as best you can. If a question is not relevant, simply answer ”N/A”.

What is your age?

Please state your gender.

What is your current profession / study ?

What is your country of origin?

How often do you use a computer?

How often do you participate in meetings?

How would you characterize your typical meetings (e.g. subject matter, goal, atmosphere)?

When you have missed a meeting, how do you typically catch up (e.g. read the minutes, ask other participants) ?

A.2 Instructions - Condition 3

Task Instructions

This browser presents you with a record of four meetings attended by four individuals. The four meetings are in a series (A,B,C,D), and the overall goal of the meetings was for the group to design a television remote control

together. The four participants are a project manager (PM), user interface designer (UI), marketing expert (ME) and an industrial designer (ID).

Using this browser, you can read the transcript of each meeting, watch the video and listen to the audio of each meeting, and you are also presented with summaries of what happened in each meeting. These summaries are divided into four sections: Decisions, Actions, Goals, and Problems. Repeatedly clicking on a sentence within a summary will take you to related sentences within the meeting transcript in turn.

We are interested in the group's decision-making ability, and therefore ask you to evaluate and summarize a particular aspect of their discussion.

The group discussed the issue of separating the commonly-used functions of the remote control from the rarely-used functions of the remote control. What was their final decision on this design issue? Please write a short summary (1-2 paragraphs) describing the final decision, any alternatives the participants considered, the reasoning for and against any alternatives (including why each was ultimately rejected), and in which meetings the relevant discussions took place. Please write your summary in the browser tab labelled Writing tab.

You have a total of 45 minutes for this task. Please leave yourself enough time to complete the written summary. I will give you a warning when there are 5 minutes remaining. Please signal me when you are ready to begin the experiment. If you finish before the allotted time, please signal me to end the experiment. Thank you very much for your time.

A.3 Post-Questionnaire - Conditions 3 and 4

For each statement in the following section, indicate how strongly you agree or disagree with the statement by providing the most relevant number (for example, 1=disagree strongly and 5=agree strongly)

1. I found the meeting browser intuitive and easy to use. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

2. I was able to find all of the information I needed. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

3. I was able to efficiently find the relevant information. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

4. I feel that I completed the task in its entirety. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

5. I understood the overall content of the meeting discussions. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

6. The task required a great deal of effort. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

7. I had to work under pressure. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

8. I had the tools necessary to complete the task efficiently. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

9. I would have liked additional information about the meetings. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

10. It was difficult to understand the content of the meetings using this browser. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

In the following section, please answer the questions with a short response of 1-3 sentences.

11. How useful did you find the meeting summaries?

12. What information would you have liked that you didn't have?

Appendix B

Cuewords Lists

This appendix provides the cueword lists for the AMI and ICSI corpora for manual and ASR transcripts, as used in Chapter 4. Each term represents a stem. Terms are listed in descending order according to ratio of frequency in extracted dialogue acts to frequency in non-extracted dialogue acts.

Rank	AMI-MAN	AMI-ASR	Rank	AMI-MAN	AMI-ASR
1	expect	expect	**36**	previous	suggest
2	found	component	**37**	look	gonna
3	component	found	**38**	overall	wanna
4	project	fairly	**39**	present	list
5	focus	agenda	**40**	we	little
6	group	focus	**41**	suggest	whole
7	research	project	**42**	cannot	were
8	meet	group	**43**	were	look
9	final	research	**44**	report	last
10	agenda	team	**45**	want	we
11	fairly	meet	**46**	little	interest
12	general	final	**47**	after	cannot
13	list	action	**48**	interest	present
14	particular	particular	**49**	point	want
15	decision	general	**50**	inform	saw
16	role	decision	**51**	relevant	could
17	consider	will	**52**	saw	after
18	response	discuss	**53**	could	inform
19	us	first	**54**	last	would
20	decide	especially	**55**	something	point
21	will	relevant	**56**	maybe	always
22	discuss	decide	**57**	probably	reason
23	first	role	**58**	able	then
24	team	us	**59**	definite	happy
25	option	report	**60**	would	bit
26	should	should	**61**	idea	better
27	mention	find	**62**	then	all
28	action	option	**63**	end	idea
29	whole	mention	**64**	help	explain
30	need	need	**65**	reason	something
31	find	previous	**66**	said	said
32	example	example	**67**	necessary	ask
33	important	response	**68**	ask	issue
34	especially	important	**69**	though	probably
35	gonna	consider	**70**	can	wonder

Table B.1: AMI Cuewords, Manual and ASR

Rank	AMI-MAN	AMI-ASR	Rank	AMI-MAN	AMI-ASR
1	focus	focus	**36**	start	mostly
2	fairly	soon	**37**	will	first
3	area	fairly	**38**	able	we
4	group	apparently	**39**	us	option
5	project	study	**40**	background	us
6	report	report	**41**	everything	turn
7	soon	group	**42**	we	find
8	decision	project	**43**	ask	summary
9	topic	finish	**44**	look	everybody
10	summary	response	**45**	want	arge
11	result	result	**46**	important	suggest
12	next	next	**47**	found	want
13	appar	decision	**48**	last	action
14	response	topic	**49**	bit	important
15	issue	general	**50**	first	slide
16	study	decide	**51**	list	everything
17	general	issue	**52**	end	last
18	finish	ask	**53**	time	gonna
19	decide	able	**54**	gonna	note
20	hurt	discuss	**55**	worry	whole
21	problem	end	**56**	were	research
22	mention	interpret	**57**	help	would
23	interpret	factor	**58**	need	idea
24	discuss	found	**59**	whole	probably
25	turn	inform	**60**	idea	need
26	inform	area	**61**	everybody	suppose
27	option	bit	**62**	factor	refer
28	relevant	problem	**63**	little	help
29	meet	meet	**64**	suppose	particular
30	find	mention	**65**	particular	were
31	suggest	background	**66**	note	will
32	action	bad	**67**	before	time
33	research	start	**68**	agenda	wrong
34	refer	look	**69**	previous	final
35	nice	nice	**70**	now	wont

Table B.2: ICSI Cuewords, Manual and ASR

Appendix C

Decision Audit Gold-Standard Items

This appendix contains the gold-standard items for the decision audit task as determined two human judges. These are the pieces of information from the four meetings that together work to satisfy the information need.

AGR standard for an agreement or decision, PRP stands for a proposal, INF stands for information from external sources, DIS stands for discussion, and REJ stands for rejection of a proposal or idea.

- **ES2008a**
- AGR: The remote control must be simple.
- AGR: The remote control must not have too many buttons.
- PRP: There should not be too many different remote controls just to watch T.V.
- PRP: They proposed to have one remote control with main functions and a separate one for special functions.
- DIS: They discussed personal experiences with remotes and preferences about simplicity. In general they do not like complicated remotes but also do not want a lot of separate remotes.
- DIS: They are unsure if the remote should only be for the T.V.
- **ES2008b**
- INF: The marketing expert presented a marketing study that mentioned an LCD-screen on the remote control.

- PRP: The user interface designer and the marketing expert suggest a sliding screen that hides the more complicated buttons.
- PRP: The simple features should stay on the main part.
- AGR: The essential features should have large buttons.
- AGR: There is a possibility to access extra features (that are possibly hidden) but they will not be as prominent as the main features.
- AGR: They decide to concentrate only on T.V. remote control.
- **ES2008c**
- PRP: The industrial designer mentioned the possibility of an LCD-screen.
- PRP: The user interface designer proposes to use Apple's iPod as a model for the remote control.
- PRP: There should be buttons for the advanced features such as brightness and contrast.
- PRP: The buttons for the advanced features could be included on the remote control or they could be on a little LCD-screen.
- DIS: The participants discussed having either an LCD-display or a menu that comes up on the T.V. screen.
- DIS: The disadvantage of an LCD-display is that it would be small and require backlighting.
- DIS: The disadvantage of a menu on the T.V. screen would be the readability of the text on a small T.V. screen that might be far away.
- PRP The user interface designer proposed to have a scroll menu on the (T.V.-) screen and to use push buttons to scroll. Advantage: The simpler chip can be used and that chip is also cheaper.
- AGR: They agreed to have a menu button with on-screen functions.
- AGR: They agreed on using push buttons instead of a scroll wheel.
- **ES2008d**

- PRP: They proposed to have five buttons.
  - Menu button
  - Volume up
  - Volume down
  - Channel up
  - Channel down
- PRP: The project manager suggests adding a seperate power button.
- AGR: They agreed on having six buttons including a seperate power button.
- (sliding screen / hatch never mentioned again)

Appendix D

Abstractive Cuewords

This appendix provides the list of abstractive cuewords used in Chapter 7. These are derived by comparing frequency in the training data abstracts to frequency in the training data extracts. The terms in Table D.1 are ranked in decreasing order according to that frequency ratio.

Rank	Cueword	Rank	Cueword
1	team	**26**	use
2	group	**27**	product
3	specialist	**28**	component
4	member	**29**	complain
5	manager	**30**	introduce
6	project	**31**	device
7	expert	**32**	drew
8	discuss	**33**	gave
9	remote	**34**	examine
10	design	**35**	she
11	industrial	**36**	meet
12	their	**37**	suggest
13	include	**38**	state
14	prototype	**39**	cost
15	interface	**40**	work
16	whether	**41**	misplace
17	feature	**42**	not
18	he	**43**	function
19	user	**44**	budget
20	present	**45**	material
21	participate	**46**	recognition
22	decide	**47**	incorporate
23	market	**48**	button
24	contain	**49**	her
25	announce	**50**	initial

Table D.1: Abstractive Cuewords

Appendix E

Meta and Non-Meta Comparison

Below are the meta and non-meta summaries for AMI meeting TS3003c, for comparison. The meta summary scored much more highly according to weighted f-score using the new extractive labels. Dialogue acts preceded by an asterisk indicate examples of meta comments that were not included in the low-level summary.

E.1 Meta Summary of TS3003c

speaker B is it possible to um program it s so uh you got on the left side uh or on the right side uh buttons for for shifting u up and shifting up and on the uh other uh uh o other side uh buttons for uh shifting uh for for the sound
speaker B weve got um the buttons we have to use the onoff sound onoff sound higher or lower um the numbers uh zero to uh uh nine um the general buttons m more general b one button for shifting up and shifting down uh channel
speaker D but if we would make um a changing channels and changing volume button on both sides that would certainly yield great options for the design of the remote
speaker D and i think a voice recognition function would not make the remote control much easier to use
∗ **speaker A** so the industrial designer and user interface designer are going to work together on this one
speaker A i personally think the lcd screen we wanna use with the extra information i think nobody has anything against it
speaker A so you have a but button on your docking station which you can push and then it starts beeping
speaker A its important that the corporate design image uh is going to be in the remote

speaker C and uh our d manufacturing department can also deliver single curved or double curved ca curved cases

speaker D this was for like an lcd screen like you would have on a on the the most advanced mobile phones

speaker D and on top of that the lcd screen would um help in making the remote control easier to use

speaker A but i think just a simple battery which you can reload on a docking station is just as good

speaker D personally i dont think that older people like to shake their remote control before they use it

speaker D cause we would have to make one w uh control which would fit in with a wooden cover and a plastic cover the more original one or the more standard one

speaker A so maybe a docking station will help them give the remote a place

∗ **speaker B** but uh is uh our uh research um about um bi large uh lcd sh uh display or uh just a small one uh we want to uh use

∗ **speaker A** uh f i think first of all we have to see uh it is possible to introduce kinetic energy in our budget i think

speaker C the the single curved so im not really sure what theyre gonna look like but i think its something like this

speaker A on which you can apply yeah remote controls on which you can apply different case covers for example

speaker A and then we can we can still use the voice recognition but maybe then for only the the channels

speaker D um i thought maybe we could just make one of those buttons on both the left and the right side

speaker A cant we make uh cant we make a remote which you can flip over and use on the same

speaker A and then you can make them with colour black and grey other colours as well

speaker A if we dont have the voice recognition it will it wont use a lot of energy to use um

speaker A uh requirements are uh teletext docking station audio signal small screen with some extras that uh button information

∗ **speaker A** if you wanna have a look at it its over there in the projects folder

speaker D um we also um asked if w they would if people would pay more for speech

recognition in a remote control
speaker A because uh maybe your hand is in the way if you have the display here
speaker A give your grandfather a new case for his remote control or whatever
speaker A just thats for left hand and right hand users
speaker A uh but maybe we have to make it a l a bit more fancy in one or ano another way
∗ **speaker D** um i heard our industrial designer talk about uh flat single and double curved

E.2 Non-Meta Summary of TS3003c

speaker B is it possible to um program it s so uh you got on the left side uh or on the right side uh buttons for for shifting u up and shifting up and on the uh other uh uh o other side uh buttons for uh shifting uh for for the sound
speaker D um well the trendwatchers i consulted advised that it b should be the remote control and the docking station should be telephoneshaped
speaker B weve got um the buttons we have to use the onoff sound onoff sound higher or lower um the numbers uh zero to uh uh nine um the general buttons m more general b one button for shifting up and shifting down uh channel
speaker B um double push push um if double click um so uh you get uh big uh subtitles for uh people uh um uh which c f uh who cant uh read small uh subtitles
speaker D um besides that we would advise um to bring two editions one with a woodlike colour and maybe feel and one with a greyblack colour
speaker D so they would prefer uh a design where the remote control just lies flat in the docking station
speaker B and and uh for uh shifting up a sen uh c ch channel or uh for um uh putting out uh sound or something you can uh just give a sign uh say um sound off
speaker A and a few points of interest in this meeting um are the conceptual specification of components uh conceptual specification of design and also trendwatching
speaker B um also we want to uh use a little d display uh for um for displaying the uh the functions of the buttons
speaker B only uh buttons uh for uh sound um for uh onoff um uh shifting u up uh sa uh ca channel or uh down shifting down
speaker D but if we would make um a changing channels and changing volume button

on both sides that would certainly yield great options for the design of the remote
speaker D um this is this image will give you a little bit of an impression about um the lookandfeel that um the remote should have
speaker B finding an attractive uh way to control uh the remote control um the uh i found some uh something about uh speech uh recognition
speaker D and ive consulted some additional trendwatch trendwatchers after the original trendwatchers return about what the the best design would be
speaker B um you can think about um uh when you lost your um remote control you can uh call it and um it gives an um sig signal
speaker D um for our um group were focusing on the people of sixty to eighty y years old this is um these factors are slightly more equal
speaker A uh requirements are uh teletext docking station audio signal small screen with some extras that uh button information
speaker B but uh is uh our uh research um about um bi large uh lcd sh uh display or uh just a small one uh we want to uh use
speaker B and a special uh button for shifting up uh and uh shifting down uh channel um its uh on place where um the thumb of of the
speaker B um almost uh e all uh remote controls uh are using a onoff button on that place
speaker B and um we can uh build in a function f which uh shows the channel or some uh which the t television is on
speaker D so you could imagine that uh the remote control will be standing up straight in the docking station
speaker D cause we would have to make one w uh control which would fit in with a wooden cover and a plastic cover the more original one or the more standard one
speaker D uh the remote control and the docking station should uh blend in in the in the room
speaker C uh for the casing uh the uh manufacturing department can deliver uh a flat casing single or double curved casing

Appendix F

Intersection and Union of Human Selected Dialogue Acts

For both the AMI and ICSI test sets, we find the union and intersection of dialogue acts selected by human annotators, and calculate the precision and recall on each of those sets. For the AMI test set, the intersection of selected dialogue acts is on average 23% as large as the union of selected sentences. This is perhaps unsurprising given our knowledge of the low inter-annotator agreement.

For the six ICSI test set meetings, the intersection of selected dialogue acts is on average less than 3% as large as the union of selected sentences. This partly reflects our finding that inter-annotator agreement is substantially lower for the ICSI data compared with the AMI data. Another reason for the much smaller percentage for the ICSI data is that three of the six ICSI test set meetings have more than three annotators, thus decreasing the chance that a given dialogue act would be selected by every annotator in those meetings. However, even for the three meetings with exactly three annotators, the intersection of human selected dialogue acts is very low.

Table F.1 provides the precision and recall scores for the AMI corpus, comparing *tf.idf* and *su.idf* metrics across manual and ASR transcripts.

	SUIDF Manual	SUIDF ASR	TFIDF Manual	TFIDF ASR
Union Precision	0.79	0.82	0.76	0.76
Union Recall	0.15	0.15	0.16	0.15
Intersection Precision	0.28	0.29	0.25	0.26
Intersection Recall	0.22	0.22	0.22	0.22

Table F.1: Precision and Recall for Union and Intersection, AMI Corpus

	SUIDF Manual	SUIDF ASR	TFIDF Manual	TFIDF ASR
Union Precision	0.61	0.67	0.63	0.74
Union Recall	0.068	0.082	0.086	0.10
Intersection Precision	0.032	0.033	0.023	0.05
Intersection Recall	0.13	0.013	0.10	0.24

Table F.2: Precision and Recall for Union and Intersection, ICSI Corpus

Table F.2 provides the precision and recall scores for the ICSI corpus, comparing *tf.idf* and *su.idf* metrics across manual and ASR transcripts.

Bibliography

Ang, J., Liu, Y., & Shriberg, E. (2005). Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. In *Proc. of ICASSP 2005, Philadelphia, USA*, pp. 874–877.

Arons, B. (1997). SpeechSkimmer: a system for interactively skimming recorded speech. *ACM Trans. Comput.-Hum. Interact.*, *4*(1), 3–38.

Barzilay, R., Collins, M., Hirschberg, J., & Whittaker, S. (2000). The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts. In *Proc. of the AAAI 2000, Austin, Texas, USA*, pp. 679–684.

Barzilay, R., & Elhadad, M. (1997). Using Lexical Chains for Summarisation. In *Proc. of ACL 1997, Madrid, Spain*, pp. 10–18.

Borko, H., & Bernier, C. (1975). *Abstracting Concepts and Methods*. Academic Press, New York.

Bush, V. (1945). As We May Think.. *The Atlantic Monthly*, *176*(1), 101–108.

Carbonell, J., & Goldstein, J. (1998). 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*, pp. 335–336.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, *22*(2), 249–254.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI Meeting Corpus: A Pre-Announcement. In *Proc. of MLMI 2005, Edinburgh, UK*, pp. 28–39.

Charniak, E., & Johnson, M. (2001). Edit detection and parsing for transcribed speech. In *Proc. of NAACL 2001, Pittsburgh, USA*, pp. 1–9.

Chen, F., & Withgott, M. (1992). The Use of Emphasis to Automatically Summarize a Spoken Discourse. In *Proc. of ICASSP 1992, San Francisco, USA*, pp. 229–232.

Chen, Y.-W., & Lin, C.-J. (2006). Combining SVMs with various feature selection strategies. In Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. (Eds.), *Feature extraction, foundations and applications*. Springer.

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

Christensen, H., Gotoh, Y., & Renals, S. (2008). A Cascaded Broadcast News Highlighter. *IEEE Transactions on Audio, Speech and Language Processing*, *0*(0), 0.

Church, K., & Gale, W. (1995). Inverse Document Frequency IDF: A Measure of Deviation from Poisson. In *Proc. of the Third Workshop on Very Large Corpora*, pp. 121–130.

Clarke, J., & Lapata, M. (2006). Constraint-based sentence compression: An integer programming approach. In *Proc. of COLING/ACL 2006*, pp. 144–151.

Craswell, N., Zaragoza, H., & Robertson, S. (2005). Microsoft Cambridge at TREC-14: Enterprise Track. In *Proc. of TREC-2005, Gaithersburg, Maryland, USA*.

Croft, W., & Harper, D. (1979). Using Probabilistic Models of Information Retrieval Without Relevance Information. *Journal of Documentation*, *35*, 285–295.

Dang, H. (2005). Overview of DUC 2005. In *Proc. of the Document Understanding Conference (DUC) 2005, Vancouver, BC, Canada*.

Daumé, H., & Marcu, D. (2005). Bayesian Summarization at DUC and a Suggestion for Extrinsic Evaluation. In *Proc. of DUC 2005, Vancouver, Canada*.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

DeJong, G. (1982). An Overview of the FRUMP System. In Lehnert, W. G., & Ringle, M. H. (Eds.), *Strategies for Natural Language Processing*, pp. 149–176. Lawrence Erlbaum Associates, Publishers.

Dielmann, A., & Renals, S. (2007). DBN Based Joint Dialogue Act Recognition of Multiparty Meetings. In *Proc. of ICASSP 2007, Honolulu, USA*, pp. 133–136.

Dilley, L., Breen, M., Bolivar, M., Kraemer, J., & Gibson, E. (2006). A Comparison of Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). In *Proc. of Interspeech 2006, Pittsburgh, USA*, pp. 317–320.

Dorr, B., Monz, C., Oard, D., Zajic, D., & Schwartz, R. (2004). Extrinsic Evaluation of Automatic Metrics for Summarization. Tech. rep. LAMP-TR-115,CAR-TR-999,CS-TR-4610,UMIACS-TR-2004-48, University of Maryland, College Park and BBN Technologies*.

Dorr, B., Monz, C., President, S., Schwartz, R., & Zajic, D. (2005). A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate?. In *ACL 2005, MTSE Workshop, Ann Arbor, USA*, pp. 1–8.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *J. ACM*, *16*(2), 264–285.

Endres-Niggemeyer, B. (1998). *Summarizing Information*. Springer, Berlin.

Foltz, P., Kintsch, W., & Landauer, T. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes*, *25*.

Fujii, Y., Kitaoka, N., & Nakagawa, S. (2007). Automatic Extraction of Cue Phrases for Important Sentences in Lecture Speech and Automatic Lecture Speech Summarization. In *Proc. of Interspeech 2007, Antwerp, Belgium*, pp. 2801–2804.

Fum, D., Guida, G., & Tasso, C. (1982). Forward and Backward Reasoning in Automatic Abstracting. In *Proc. of the (COLING '82), Prague, Czech Republic*, pp. 83–88.

Galley, M. (2006). A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance. In *Proc. of EMNLP 2006, Sydney, Australia*, pp. 364–372.

Girgensohn, A., Boreczky, J., & Wilcox, L. (2001). Keyframe-Based User Interfaces for Digital Video. *IEEE Computer*, *34*(9), 61–67.

Gong, Y., & Liu, X. (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA*, pp. 19–25.

Hachey, B., Murray, G., & Reitter, D. (2005). The Embra System at DUC 2005: Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space. In *Proc. of the Document Understanding Conference (DUC) 2005, Vancouver, BC, Canada*.

Hain, T., Burget, L., Dines, J., Garau, G., Wan, V., Karafiat, M., Vepa, J., & Lincoln, M. (2007). The AMI System for Transcription of Speech in Meetings. In *Proc. of ICASSP 2007,*, pp. 357–360.

Harman, D. (1992). Overview of the First Text REtrieval Conference. In *Proc. of TREC 1992, Gaithersburg, MD, USA*, pp. 1–20.

Harman, D., & Over, P. (2004). Document Understanding Conference 2004. In *Proc. of the DUC 2004, Boston, USA*.

Hersh, W., & Over, P. (2001). Interactivity at the Text Retrieval Conference (TREC). *Information Processing Management*, *37*(3), 365–367.

Hirschberg, J., Bacchiani, M., Hindle, D., Isenhour, P., Rosenberg, A., Stark, L., Stead, L., Whittaker, S., & Zamchick, G. (2001). SCANMail: Browsing and Searching Speech Data by Content. In *Proc. of Interspeech 2001, Aalborg, Denmark*, pp. 1299–1302.

Hirschberg, J., & Nakatani, C. (1998). Acoustic Indicators of Topic Segmentation. In *Proc. of ICSLP 1998, Sydney, Australia*, pp. 976–979.

Hirschberg, J., Whittaker, S., Hindle, D., Pereira, F., & Singhal, A. (1999). Finding information in audio: A new paradigm for audio browsing and retrieval. In *Proc. of the ESCA ETRW Workshop, Cambridge UK*, pp. 117–122.

Hirschman, L., Light, M., & Breck, E. (1999). Deep Read: A Reading Comprehension System. In *Proc. of ACL 1999, College Park, MD, USA*, pp. 325–332.

Hori, C., & Furui, S. (2000). Automatic Speech Summarization Based on Word Significance and Linguistic Likelihood. In *Proc. of ICASSP 2000, Istanbul, Turkey*, pp. 1579–1582.

Hori, C., & Furui, S. (2004). Speech summarization: An approach through word extraction and a method for evaluation. *IEICE Transactions on Information and Systems*, *E87-D(1)*, 15–25.

Hori, C., Furui, S., Malkin, R., Yu, H., & Waibel, A. (2002). Automatic speech summarization applied to English broadcast news speech. In *Proc. of the ICASSP 2002, Orlando, USA*, pp. 9–12.

Hori, T., Hori, C., & Minami, Y. (2003). Speech summarization using weighted finite-state transducers. In *Proc. of Interspeech 2003, Geneva, Switzerland*, pp. 2817–2820.

Hovy, E., & Lin, C.-Y. (1999). Automated Text Summarization in SUMMARIST. In Mani, I., & Maybury, M. T. (Eds.), *Advances in Automatic Text Summarization*, pp. 81–94. MITP.

Hovy, E., Lin, C.-Y., & Zhou, L. (2005). A BE-Based Multi-Document Summarizer with Query Interpretation. In *Proc. of DUC 2005, Vancouver, CA*.

Hovy, E., Lin, C.-Y., Zhou, L., & Fukumoto, J. (2006). Automated Summarization Evaluation with Basic Elements. In *Proc. of LREC 2006, Genoa, Italy*.

Hsueh, P.-Y., Kilgour, J., Carletta, J., Moore, J., & Renals, S. (2007). Automatic Decision Detection in Meeting Speech. In *Proc. of MLMI 2007, Brno, Czech Republic*.

Hsueh, P.-Y., & Moore, J. (2006). Acoustic Topic Segmentation and Labeling in Multiparty Dialogue. In *Proc. of IEEE Spoken Language Technology Workshop, Aruba*, pp. 98–101.

Huang, J., Zweig, G., & Padmanabhan, M. (2001). Information Extraction from Voicemail. In *Proc. of ACL 2001, Toulouse, France*, pp. 290–297.

Jacobs, P., & Rau, L. (1990). SCISOR: Extracting Information from On-Line News. *CACM*, *33*(11), 88–97.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., & Wooters, C. (2003). The ICSI Meeting Corpus. In *Proc. of IEEE ICASSP 2003, Hong Kong, China*, pp. 364–367.

Janos, J. (1979). Theory of functional sentence perspective and its application for the purposes of automatic extracting. *Information Processing Management*, *15*(1), 19–25.

Jansche, M., & Abney, S. (2002). Information Extraction from Voicemail Transcripts. In *Proc. of EMNLP 2002, Philadelphia, USA*, pp. 320–327.

Jing, H., Barzilay, R., McKeown, K., & Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. In *Proc. of the AAAI Symposium on Intelligent Summarization, Stanford, USA*, pp. 60–68.

Jones, K. S. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrievall. *Journal of Documentation*, *28*, 11–21.

Jones, K. S. (1999). Automatic Summarizing: Factors and Directions. In Mani, I., & Maybury, M. (Eds.), *Advances in Automatic Text Summarization*, pp. 1–12. MITP.

Jones, K. S., & Galliers, J. (1995). *Evaluating Natural Language Processing Systems: An Analysis and Review*. No. 1083 in Lecture Notes in Artificial Intelligence. Springer.

Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing Management*, *36*(6), 779–808.

Kameyama, M., & Arima, I. (1994). Coping with aboutness complexity in information extraction from spoken dialogues. In *Proc. of ICSLP 1994, Yokohama, Japan*, pp. 87–90.

Kimber, D., Wilcox, L., Chen, F., & Moran, T. (1995). Speaker segmentation for browsing recorded audio. In *Proc. of CHI 95, Denver, United States*, pp. 212–213.

Kleinbauer, T., Becker, S., & Becker, T. (2007). Combining Multiple Information Layers for the Automatic Generation of Indicative Meeting Abstracts. In *Proc. of ENLG 2007, Dagstuhl, Germany*.

Knight, K., & Marcu, D. (2000). Statistics-Based Summarization - Step One: Sentence Compression. In *Proc. of AAAI 2000, Austin, Texas, USA*, pp. 703–710.

Kolluru, B., Gotoh, Y., & Christensen, H. (2005). Multi-Stage Compaction Approach to Broadcast News Summarisation. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pp. 69–72.

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

Kraaij, W., & Post, W. (2006). Task based evaluation of exploratory search systems. In *Proc. of SIGIR 2006 Workshop, Evaluation Exploratory Search Systems, Seattle, USA*, pp. 24–27.

Kupiec, J., Pederson, J., & Chen, F. (1995). A Trainable Document Summarizer. In *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA*, pp. 68–73.

Lacatusu, F., Hickl, A., Aarseth, P., & Taylor, L. (2005). Lite-GISTexter at DUC 2005. In *Proc. of DUC 2005, Vancouver, CA*.

Lee, D., Erol, B., Graham, J., Hull, J., & Murata, N. (2002). Portable Meeting Recorder. In *Proc. of ACM Multimedia 2002, Juan Les Pins, France*, pp. 493–502.

Lin, C.-Y. (2004). Looking for a Few Good Metrics: Automatic Summarization Evaluation - How Many Samples Are Enough. In *Proc. of NTCIR 2004, Tokyo, Japan*, pp. 1765–1776.

Lin, C.-Y., & Hovy, E. H. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proc. of HLT-NAACL 2003, Edmonton, Calgary, Canada*, pp. 71–78.

Lin, J., & Demner-Fushman, D. (2005). Evaluating Summaries and Answers: Two Sides of the Same Coin?. In *Proceedings of the ACL 2005 MTSE Workshop, Ann Arbor, USA*, pp. 41–48.

Lin, J., & Demner-Fushman, D. (2006). Will pyramids built of nuggets topple over?. In *Proc. of HLT/NAACL 2006*, pp. 383–390 Morristown, NJ, USA. Association for Computational Linguistics.

Liu, Y., Liu, F., Li, B., & Xie, S. (2007). Do Disfluencies Affect Meeting Summarization: A Pilot Study on the Impact of Disfluencies. In *Proc. of MLMI 2007, Brno, Czech Republic*, p. poster.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, *2*(2), 159–165.

Mani, I. (2001a). *Automatic Summarization*. John Benjamin, Amsterdam, NL.

Mani, I. (2001b). Summarization Evaluation: An Overview. In *Proc. of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, Tokyo, Japan*, pp. 77–85.

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., & Sundheim, B. (1999). The TIPSTER SUMMAC Text Summarization Evaluation. In *Proc. of EACL 1999, Bergen, Norway*, pp. 77–85.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281.

Marcu, D. (1995). Discourse Trees Are Good Indicators of Importance in Text. In Mani, I., & Maybury, M. T. (Eds.), *Advances in Automatic Text Summarization*, pp. 123–136. MITP, Cambridge, MA.

Marcu, D. (1997). From Discourse Structures to Text Summaries. In *Proc. of ACL 1997 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain*, pp. 82–88.

Maron, M. E., & Kuhns, J. L. (1960). On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, *7*(3), 216–244.

Maskey, S., & Hirschberg, J. (2005). Comparing Lexial, Acoustic/Prosodic, Discourse and Structural Features for Speech Summarization. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pp. 621–624.

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

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

McDonald, R. (2006). Discriminative Sentence Compression with Soft Syntactic Evidence. In *Proc. of EACL 2006, Trento, Italy*, pp. 297–304.

McKeown, K., Hirschberg, J., Galley, M., & Maskey, S. (2005). From Text to Speech Summarization. In *Proc. of ICASSP 2005, Philadelphia, USA*, pp. 997–1000.

Mori, T. (2002). Information Gain Ratio as Term Weight: The Case of Summarization of IR results. In *Proc. of COLING 2002, Taipei, Taiwan*, pp. 688–694.

Morris, A., Kasper, G., & Adams, D. (1992). The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance. *Information Systems Research*, *3*(1), 17–35.

Murray, G., & Renals, S. (2007). Term-Weighting for Summarization of Multi-Party Spoken Dialogues. In *Proc. of MLMI 2007, Brno, Czech Republic*, pp. 155–166.

Murray, G., Renals, S., & Carletta, J. (2005a). Extractive Summarization of Meeting Recordings. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pp. 593–596.

Murray, G., Renals, S., Carletta, J., & Moore, J. (2005b). Evaluating Automatic Summaries of Meeting Recordings. In *Proc. of the ACL 2005 MTSE Workshop, Ann Arbor, MI, USA*, pp. 33–40.

Murray, G., Renals, S., Moore, J., & Carletta, J. (2006). Incorporating Speaker and Discourse Features into Speech Summarization. In *Proc. of the HLT-NAACL 2006, New York City, USA*, pp. 367–374.

Nakatani, C., & Hirschberg, J. (1993). A speech-first model for repair detection and correction. In *Proc. of ACL 1993, Columbus, Ohio, USA*, pp. 46–53.

Nenkova, A., & Passonneau, B. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. In *Proc. of HLT-NAACL 2004, Boston, MA, USA*, pp. 145–152.

Nenkova, A., Passonneau, R., & McKeown, K. (2007). The Pyramid Method: Incorporating Human Content Selection Variatin in Summarization Evaluation. *ACM Transactions on Computational Logic*, *4*(2), 1–23.

Ohtake, K., Yamamoto, K., Toma, Y., Sado, S., Masuyama, S., & Nakagawa, S. (2003). 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*, pp. 167–170.

Ono, K., Sumita, K., & Miike, S. (1994). Abstract Generation Based on Rhetorical Structure Extraction. In *Proc. of COLING 1994, Kyoto, Japan*, pp. 344–348.

Orasan, C., Pekar, V., & Hasler, L. (2007). A Comparison of Summarisation Methods Based on Term Specificity Estimation. In *Proc. of LREC 2004, Lisbon, Portugal*, pp. 1037–041.

Paice, C. (1980). The Automatic Generation of Literary Abstracts: An Approach Based on Identification of Self-Indicating Phrases. In Norman, O., Robertson., S., van Rijsbergen, C., & Williams, P. (Eds.), *Information Retrieval Research*, pp. 172–191. Butterworth, London.

Papineni, K. (2001). Why Inverse Document Frequency?. In *Proc. of NAACL 2001*, pp. 1–8.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2001). Bleu: a method for automatic evaluation of machine translation..

Pollock, J., & Zamora, A. (1975). Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Computer Sciences*, *15*(4), 226–232.

Porter, M. (1997). An algorithm for suffix stripping. In *Readings in information retrieval*, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Radev, D., Blair-Goldensohn, S., & Zhang, Z. (2001). Experiments in single and multi-document summarization using MEAD. In *Proc. of DUC 2001, New Orleans, LA, USA*.

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

Radev, D., & Tam, D. (2003). Summarization Evaluation Using Relative Utility. In *Proc. of CIKM 2003, New Orleans, USA*, pp. 508–511.

Rath, G., Resnick, A., & Savage, R. (1961). The Formation of Abstracts by the Selection of Sentences: Part 1: Sentence Selection by Man and Machines. *American Documentation*, *12*(2), 139–141.

Reithinger, N., Kipp, M., Engel, R., & Alexandersson, J. (2000). Summarizing multilingual spoken negotiation dialogues. In *Proc. of ACL 2000, Hong Kong*, pp. 310–317 Morristown, NJ, USA. Association for Computational Linguistics.

Rennie, J., & Jaakkola, T. (2005). Using Term Informativeness for Named Entity Recognition. In *Proc. of SIGIR 2005, Salvador, Brazil*, pp. 353–360.

Rijsbergen, C. V. (1979). *Information Retrieval, 2nd Ed.* Butterworths, London, UK.

Robertson, S. (2004). Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation*, *60*, 503–520.

Robertson, S., & Jones, K. S. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, *35*, 129–146.

Robertson, S., & Jones, K. S. (1994). Simple, Proven Approaches to Text Retrieval. *University of Cambridge Computer Laboratory Technical Report TR-356*, *356*.

Robertson, S., Walker, S., & Hancock-Beaulieu, M. (1998). Okapi at TREC-7. In *Proc. of TREC 1998, Gaithersburg, USA*, pp. 253–264.

Rohlicek, J. R. (1992). Gisting continuous speech. In *Proc. of ICASSP 1992, San Francisco, USA*, pp. 384–384.

Rush, J., Zamora, A., & Salvador, R. (1971). Automatic Abstracting and Indexing. II, Production of Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria. *JASIS*, *22*(4), 260–274.

Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, *24*, 513–523.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, NY, NY, USA.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., , & Carvey, H. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of SIGdial Workshop on Discourse and Dialogue, Cambridge, MA, USA*, pp. 97–100.

Shriberg, E., & Stolcke, A. (2004). Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing. In *Proc. of International Conference on Speech Prosody 2004, Nara, Japan*.

Shriberg, E., Stolcke, A., & Baron, D. (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proc. of Interspeech 2001, Aalborg, Denmark*, pp. 1359–1362.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A Standard for Labeling English Prosody. In *Proc. of ICSLP 1992, Banff, Canada*, pp. 867–870.

Simpson, S., & Gotoh, Y. (2005). Towards Speaker Independent Features for Information Extraction from Meeting Audio Data. In *Proc. of MLMI 2005, Edinburgh, UK*, p. poster.

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *24*(4), 35–43.

Song, Y., Han, K., & Rim, H. (2004). A Term Weighting Method Based on Lexical Chain for Automatic Summarization. In *Proc. of CICLing 2004, Seoul, Korea*, pp. 636–639.

Steedman, M. (2000). *The syntactic process*. MIT Press, Cambridge, MA, USA.

Steedman, M. (2007). Information-Structural Semantics for English Intonation. In Lee, C., Gordon, M., & Büring, D. (Eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, pp. 245–264. Springer.

Steinberger, J., & Ježek, K. (2004). Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. In *Proc. of ISIM 2004, Roznov pod Radhostem, Czech Republic*, pp. 93–100.

T. Kikuchi, S. F., & Hori, C. (2003). Two-stage Automatic Speech Summarization by Sentence Extraction and Compaction. In *Proce. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition 2003, Tokyo, Japan*, pp. 207–210.

Teufel, S., & Moens, M. (1997). Sentence Extraction as a Classification Task. In *Proc. of ACL 1997, Workshop on Intelligent and Scalable Text Summarization, Madrid, Spain*, pp. 58–65.

Teufel, S., & Moens, M. (1999). Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. In Mani, I., & Maybury, M. (Eds.), *Advances in Automatic Text Summarization*, pp. 155–171. MITP.

Teufel, S., & van Halteren, H. (2004). Evaluating Information Content by Factoid Analysis: Human Annotation and Stability. In *Proc. of EMNLP 2004, Barcelona, Spain*, pp. 419–426.

Tucker, S., & Whittaker, S. (2004). Accessing Multimodal Meeting Data: Systems, Problems and Possibilities. In *Proc. of MLMI 2004, Martigny, Switzerland*, pp. 1–11.

Tucker, S., & Whittaker, S. (2006). Time is of the essence: an evaluation of temporal compression algorithms. In *Proc. of SIGCHI 2006, Montreal, Canada*, pp. 329–338.

Valenza, R., Robinson, T., Hickey, M., & Tucker, R. (1999). Summarization of spoken audio through information extraction. In *Proc. of the ESCA Workshop on Accessing Information in Spoken Audio, Cambridge UK*, pp. 111–116.

Voorhees, E. (2004). Overview of the TREC 2004 Question Answering Track. In *Proc. of TREC 2004*, pp. 83–105.

Voorhees, E., & Harman, D. (1999). Overview of the Seventh Text REtrieval Conference (TREC-7). In *Proc. of TREC 1999*, pp. 1–24.

Waibel, A., Bett, M., Finke, M., & Stiefelhagen, R. (1998). Meeting browser: Tracking and summarizing meetings. In Penrose, D. E. M. (Ed.), *Proc. of the Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, USA*, pp. 281–286.

Wellner, P., Flynn, M., & Guillemot, M. (2004). Browsing Recorded Meetings with Ferret. In *Proc. of MLMI 2004, Martigny, Switzerland*, pp. 12–21.

Wellner, P., Flynn, M., Tucker, S., & Whittaker, S. (2005). A Meeting Browser Evaluation Test. In *Proc. of the SIGCHI Conference on Human Factors in Computing*

Systems 2005, Portland, OR, USA, pp. 2021–2024 New York, NY, USA. ACM Press.

Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick, G., & Rosenberg, A. (2002). SCANMail: a voicemail interface that makes speech browsable, readable and searchable. In *Proc. of the SIGCHI 2002, Minneapolis, Minnesota, USA*, pp. 275–282 New York, NY, USA. ACM.

Whittaker, S., Tucker, S., Swampillai, K., & Laban, R. (2008). Design and Evaluation of Systems to Support Interaction Capture and Retrieval. *Personal and Ubiquitous Computing*, *0*(0), 0.

Widdows, D., Dorow, B., & Chan, C.-K. (2003). Using Parallel Corpora to Enrich Multilingual Lexical Resources. In *Proc. of LREC 2003, Las Palmas, Canary Islands*, pp. 240–245.

Zechner, K. (2002). Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Computational Linguistics*, *28*(4), 447–485.

Zechner, K., & Waibel, A. (2000). Minimizing Word Error Rate in Textual Summaries of Spoken Language. In *Proc. of NAACL 2000, Seattle, WA, USA*, pp. 186–193.

Zhang, J., Chan, H., Fung, P., & Cao, L. (2007). Comparative Study on Speech Summarization of Broadcast News and Lecture Speech. In *Proc. of Interspeech 2007, Antwerp, Belgium*, pp. 2781–2784.

Zhu, X., & Penn, G. (2006). Summarization of Spontaneous Conversations. In *Proc. of Interspeech 2006, Pittsburgh, USA*, pp. 1531–1534.

Zipf, G. (1935). *Psycho-Biology of Languages*. Houghton-Mifflin, Boston, USA.