Meeting Decision Detection: Multimodal Information Fusion for Multi-Party Dialogue Understanding

Pei-Yun Sabrina Hsueh

Doctor of Philosophy

Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh

2009

Abstract

Modern advances in multimedia and storage technologies have led to huge archives of human conversations in widely ranging areas. These archives offer a wealth of information in the organization contexts. However, retrieving and managing information in these archives is a time-consuming and labor-intensive task. Previous research applied keyword and computer vision-based methods to do this. However, spontaneous conversations, complex in the use of multimodal cues and intricate in the interactions between multiple speakers, have posed new challenges to these methods. We need new techniques that can leverage the information hidden in multiple communication modalities – including not just "what" the speakers say but also "how" they express themselves and interact with others.

In responding to this need, the thesis inquires into the multimodal nature of meeting dialogues and computational means to retrieve and manage the recorded meeting information. In particular, this thesis develops the Meeting Decision Detector (MDD) to detect and track decisions, one of the most important outcomes of the meetings. The MDD involves not only the generation of extractive summaries pertaining to the decisions ("decision detection"), but also the organization of a continuous stream of meeting speech into locally coherent segments ("discourse segmentation").

This inquiry starts with a corpus analysis which constitutes a comprehensive empirical study of the decision-indicative and segment-signalling cues in the meeting corpora. These cues are uncovered from a variety of communication modalities, including the words spoken, gesture and head movements, pitch and energy level, rate of speech, pauses, and use of subjective terms. While some of the cues match the previous findings of speech segmentation, some others have not been studied before.

The analysis also provides empirical grounding for computing features and integrating them into a computational model. To handle the high-dimensional multimodal feature space in the meeting domain, this thesis compares empirically feature discriminability and feature pattern finding criteria. As the different knowledge sources are expected to capture different types of features, the thesis also experiments with methods that can harness synergy between the multiple knowledge sources.

The problem formalization and the modeling algorithm so far correspond to an optimal setting: an off-line, post-meeting analysis scenario. However, ultimately the MDD is expected to be operated online – right after a meeting, or when a meeting is still in progress. Thus this thesis also explores techniques that help relax the optimal setting, especially those using only features that can be generated with a higher

degree of automation. Empirically motivated experiments are designed to handle the corresponding performance degradation.

Finally, with the users in mind, this thesis evaluates the use of query-focused summaries in a decision debriefing task, which is common in the organization context. The decision-focused extracts (which represent compressions of 1%) is compared against the general-purpose extractive summaries (which represent compressions of 10-40%). To examine the effect of model automation on the debriefing task, this evaluation experiments with three versions of decision-focused extracts, each relaxing one manual annotation constraint. Task performance is measured in actual task effectiveness, user-generated report quality, and user-perceived success. The users' clicking behaviors are also recorded and analyzed to understand how the users leverage the different versions of extractive summaries to produce abstractive summaries.

The analysis framework and computational means developed in this work is expected to be useful for the creation of other dialogue understanding applications, especially those that require to uncover the implicit semantics of meeting dialogues.

Acknowledgements

When I was fifteen, I decided that I would like to become a writer. Fifteen years later, here I am writing the acknowledgement of my first book – my thesis. Although this book is not exactly the type I envisioned and probably is not going to be read by a large audience, I have made my first step towards what I have intended to do all along.

First, I would like to take the opportunity to thank my advisor, Prof. Johanna Moore, for her guidance throughout my PhD years, which has prepared me for my academic career. In my first year in Edinburgh, we talked about the first draft of my PhD proposal, which was of course way too ambitious. In a recent discussion with her, I found out that she still remembered what I claimed to do in the proposal to date. Compared to then, now I have the means to set realistic goals, which strucked me that I have really come a long way in these years. It is a key to a magic world. Yet I did not lose the driving force behind the effort. I owed a great deal to Johanna for this.

Many thanks should also go to my second supervisor, Prof. Steve Renals, my thesis committee (Prof. Steve Whittaker, Dr. Simon King, and Dr. Tilman Becker), my first year report committee (Dr. Alex Lascarides) and my DDD committee (Dr. Mirella Lapata). To my thesis committee: thanks for coming all the way to Edinburgh on such a stormy day. Your feedbacks and encouragement have been of great importance to me.

I also owe a lot of thanks to the whole AMI/AMIDA team at German Research Centre for Artificial Intelligence (DFKI), Netherlands Organisation for Applied Scientific Research (TNO), University of Twente, Brno University of Technology in Czech Republic, Idiap Research Institute, University of Sheffield, Munich University of Technology, and ICSI Berkeley, for their continuous support. I cannot possibly name everyone who has helped me. Jean Carletta, Theresa Wilson, Gabriel Murray, Jonathan Kilgour, Mike Lincoln, Alfred Dielmann, Weiqun Xu, Thomas Kleinbauer, Simon Tucker, Wessel Kraaij, StephanRaaijmakers, Wilfred Post, Anton Nijholt, Natasa Jovanovic, Dennis Riedsma, Petr Schwarz, Igor Szoke, Honza Cernocky, Phil Green, and the AMI-ASR team.

We have also had the most wonderful administrative support from our Edinburgh staff. David, Avril, Jenny, you are the best.

Over the years, I have had the privilege to attend conferences and workshops under the funding opportunities from the AMI and AMIDA project. This is really important to my career development. The Google Anita Borg Memorial Scholarship program

has also come as an eye-opening experience. Those amazing female scientists and engineers I met along the way truly inspired me.

My colleagues at the University of Edinburgh have also been of great help. We have the most wonderful team at the Centre for Speech Technology Research (CSTR), the Human Communication Research Centre (HCRC), the Institute for Com- municating and Collaborative Systems (ICCS) and the Department of Linguistics and English Language. I am proud to have been part of it and contributed to its growth. We have had many inspiring discussion in the machine learning discussion group, the dialogue system discussion group. I personally have also benefited a lot from my internal support group (Heriberto, Ivan, Zhang) and my dearest Buccleuch Place gang (Markus, Chris, Andrew, David T., Andi, David D., Sebastian, James, Abi, Ruken, Ben, Alex, Verena, Vera, Lexi, Sam, Michael, XingLong, SongFang, XieBei, and so many new friends.). The fond memory of Buccleuch Place will stay with me for the rest of my life.

Finally, this thesis would not have been possible without support from my family. Thanks a lot for putting up with my endless years in school. It is time to get a real job now, I know. Steven, thanks for being with me through the process, both when it was exciting and when it was depressing. Your optimism has been unquenchable, and your constructive suggestions have always been taken to heart.

In “Alice in the Wonderland”, when Alice felt lost in the woods on her way chasing the rabbit, she asked the Cheshire Cat for directions. “Oh, you’re sure to do that”, said the Cat, “If you only walk long enough.” I would like to close my acknolwedgement with this quote. Research is a long and windy journey, but as long as we explore and experience, we will eventually find our own directions.

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.

(*Pei-Yun Sabrina Hsueh*)

Table of Contents

1 Introduction **1**

1.1 Background 1

1.2 In Search of Decisions in Meeting Speech: Application Needs 4

1.3 Research Questions 6

1.3.1 What techniques can be adapted to structure the multimedia archives of meeting speech? 8

1.3.2 What multimodal and multiparty cues can be used to identify decision discussions? 9

1.3.3 How to integrate information that comes in from multiple communication modalities? 10

1.4 Meeting Decision Detector (MDD): Task Overview 12

1.4.1 Decision detection: Generating decision-focused excerpt . . . 13

1.4.2 Discourse segmentation: Determining relevant contexts 15

1.5 Goals, Future, and Guide to Remaining Chapters 16

2 Meeting Dialogue Understanding **18**

2.1 Introduction 18

2.2 Processing Meeting Dialogue: the Basics 18

2.3 Decision Making Process Modeling 20

2.4 Conversation Modeling 24

2.4.1 Modeling what the speakers say and mean 24

2.4.2 Speech acts and collaborative plans 26

2.4.3 Modeling argument intent and structure 27

2.4.4 Features Characterizing Meeting Conversations 28

2.5 Toward Automatic Derivation of Discourse Structure 30

2.5.1 Semantic Clustering 32

2.5.2 Sequence decoding 34

2.5.3 Feature-based classification . . . 35
2.6 Meeting Dialogue Processing . . . 36
2.6.1 Hierarchical segmentation . . . 37
2.6.2 Dialogue act classification . . . 39
2.6.3 Argument intent recovery . . . 40
2.6.4 Meeting affect detection . . . 42
2.6.5 Summarization . . . 44
2.7 Summary . . . 45

3 Corpus and Annotation **47**
3.1 Meeting Corpus . . . 47
3.1.1 Corpora used: the ICSI and AMI meeting corpus . . . 49
3.1.2 Transcription and ASR . . . 50
3.2 Multi-Layer Annotation . . . 51
3.2.1 Hierarchical discourse segmentation . . . 51
3.2.2 Dialogue act class labeling . . . 58
3.2.3 Extractive and abstractive summarisation . . . 59

4 Towards Shallow Processing of Meeting Speech **63**
4.1 Introduction . . . 63
4.2 Analysis of Lexical Cues . . . 63
4.3 Analysis of Prosodic Patterns . . . 67
4.4 Analysis of Dialogue Context and Dialogue Acts . . . 70
4.4.1 AMI-specific context . . . 70
4.4.2 Decision-indicative dialogue act type . . . 71
4.5 Analysis of Subjective Cues . . . 72
4.6 Summary . . . 76

5 Meeting Decision Detection **78**
5.1 Introduction . . . 78
5.2 Related Work . . . 83
5.3 Methodology . . . 83
5.3.1 The Maximum Entropy approach . . . 84
5.3.2 Data . . . 87
5.3.3 Feature extraction . . . 88
5.3.4 Multimodal feature integration . . . 93

5.4 Experiment 1: Decision Detection from Extractive Summaries 97
5.4.1 Decision-related dialogue act detection 98
5.4.2 Decision-related discourse segment detection 100
5.4.3 Effects of combining lexical features with other feature classes 101
5.5 Experiment 2: Detecting Decisions from Complete Recordings 102
5.5.1 Effects of combining non-lexical feature classes 104
5.6 Experiment 3: Exploring Automatically Generated Features 105
5.6.1 Using automatically generated DA-class features 105
5.6.2 Using automatically generated words 106
5.7 Experiment 4: Exploring the Use of Subjective Term Features 108
5.8 Experiment 5: Multimodal Integration as Feature Selection 110
5.8.1 Comparing feature discriminability measures 110
5.8.2 Comparing lexical and multimodal feature selection methods . 111
5.9 Experiment 6: Multimodal Integration as Ensemble Modeling 113
5.10 Discussion 115
5.10.1 Can decision-related conversations be detected automatically? 117
5.10.2 Can the decision detection component be operated fully automatically? 119
5.10.3 How to integrate multiple knowledge sources effectively? . . 120
5.11 Summary and Limitations 121

6 Meeting Discourse Segmentation: Determining Relevant Contexts 124
6.1 Introduction 124
6.2 Related Work 128
6.2.1 Unsupervised semantic clustering 129
6.2.2 Audio-video features beyond words 129
6.2.3 Supervised feature-based classification 130
6.2.4 Moving from off-line to on-line scenario 130
6.3 Methodology 132
6.3.1 Data 132
6.3.2 Evaluation metrics 133
6.4 ICSI Meeting Segmentation 135
6.4.1 Experiment 1: Off-line ICSI discourse segmentation from human transcripts 138
6.4.2 Experiment 2: Effects of statistically learned cue phrases . . . 140

6.4.3 Experiment 3: Online segmentation from ASR transcripts . . 142
6.5 From ICSI to AMI Meeting Segmentation 144
6.5.1 Using multimodal and multiparty interaction features 146
6.5.2 Using phonetic transcription 147
6.5.3 Using speaker activity information 148
6.6 AMI Meeting Segmentation 149
6.6.1 Experiment 4: Off-line AMI discourse segmentation from human transcripts 149
6.6.2 Experiment 5: Segmentation from ASR transcripts 152
6.6.3 Experiment 6: Segmenting directly over audio signals 154
6.7 Discussion 157
6.7.1 How to adapt the methods previously developed in text and broadcast news segmentation to segment meetings? 157
6.7.2 What are the effective knowledge sources serving to find discourse segments of different types? 160
6.7.3 How to adapt the offline segmenters to be operated online or right after the end of a meeting? 162
6.8 Summary and Limitation 163

7 Task-Oriented Evaluation of Meeting Decision Detector 167
7.1 Introduction 167
7.2 Related Work 168
7.2.1 Extractive Summarization: General v.s. Query-Driven Approach 168
7.2.2 Extractive summary evaluation 169
7.2.3 Extrinsic evaluation 170
7.3 Methodology 171
7.3.1 Task overview 171
7.3.2 Meeting Corpus and Gold Standard 173
7.3.3 Meeting Browser Interface 174
7.3.4 Manual decision-focused extracts 175
7.3.5 Automatic decision-focused extracts 175
7.3.6 Automatic general-purpose extracts 176
7.3.7 Automatic speech recognized transcription 177
7.3.8 Experiment design 177
7.4 Results 181

7.4.1 Main Effect of Summary Display Type on Decision Debriefing 182
7.4.2 Pairwise Comparison 185
7.5 Discussion 188
7.5.1 Use of audio-video aids 189
7.5.2 Use of decision-focused extracts directly on decision debriefing 191
7.6 Conclusion 191

8 Conclusion **198**
8.1 The Corpus Analysis of Decision-indicative and Discourse Segment-signalling Cues in Meeting Dialogues 199
8.1.1 The development of the lexical and multimodal feature selection framework for empirical analysis 199
8.1.2 The identification of key properties of decision-related discussions 201
8.1.3 The identification of key properties of discourse segment boundaries 201
8.2 The Automatic Derivation of Decision-related Meeting Discussions . 203
8.2.1 The integration of the lexical, multimodal, and multiparty information 203
8.2.2 Towards online processing 204
8.2.3 The advanced development of the feature selection framework 205
8.2.4 The advanced development of the knowledge source integration framework 206
8.2.5 The task-based evaluation of decision-focused summary displays 207
8.3 Limitation and Future Direction 208
8.3.1 The liability of decision-based meeting summarization 208
8.3.2 Automatic detection of argument process and outcome 208
8.3.3 The Development of Meeting Dialogue Understanding Applications 210

A Appendix A: Pre-questionnaire **213**

Bibliography **233**

List of Figures

1.1 *Illustration of the increasing importance of online videos.* 2

1.2 *Example audio-visual archives of various types of speech. In the clockwise order from the lower left corner are SciVee, Blinkx, YouTube, and an internal research seminar.* 3

1.3 *Example meeting browser augmented with the plug-in of meeting excerpts. The bottom right plug-in displays a set of meeting sequences that are representative of group design decisions. When users click on any one of these selected sequences, the focus of the meeting transcript shown on the bottom left will be switched to where the sequence of interest is located.* 6

1.4 *Example meeting browser that is augmented with the plug-in that displays human-written meeting summaries. The bottom right plug-in displays the abstracts that include the decisions made in the meeting. When users click on any one of the decisions in the abstract, the focus of the meeting transcript will be switched to where the participants are discussing the selected decision.* 7

1.5 *A list of major discussions (-) and sub-segments (-+) in an example meeting. The number of asterisks (*) indicates the number of decisions within each discussion segment.* 11

1.6 *Example excerpt of the decision made in the discussion about "how to find (the product) when misplaced".* 12

1.7 *Steps involved in the Meeting Decision Dection system.* 14

2.1 *Possible actions in the IBIS model. (Adapted from Kunz and Ritte (1970).)* . 29

2.2 *CALO ontology. Events are depicted using ovals, entities using rectangles, relations using arrows.* 41

2.3 *Example argument diagram based on the IBIS model.* 42

2.4 *Example TAS argument diagram.* 43

3.1 *The NXT tool is used to facilitate the multimodal annotation work. It contains built-in tools for media sync and data analysis and allows other customized plug-ins. In this example, the plug-in on the right hand side shows the topic structure of the meeting.* 52
3.2 *A list of major discussions (-) and sub-segments (-+) in the meeting Bed003 of the ICSI corpus.* 53
3.3 *Three-phase procedure for annotating decision-making DAs in the AMI corpus: abstractive summarization, extractive summarization, and decision linking.* 60

4.1 *Role distribution of speakers in general discussions (left) and decision speakers (right).* 70
4.2 *Meeting type distribution of speakers in general discussions (left) and decision speakers (right).* 71
4.3 *Frequency of dialogue act types in key decision-related discussions.* 72

5.1 *Example application that demonstrates the use of decision DS information. The bottom right component shows a list of discourse segments in an example meeting. The discourse segments shaded in red are those that contain at least one decision. The number shown in parentheses following each segment label indicates the number of decisions reached within the segment.* 79
5.2 *Example excerpt composed of decision DAs. The number in the parenthesis indicates the position of the selected DA in a discourse segment. In this example, annotators have selected four dialogue acts to represent the design decision of "how to find (the remote) when misplaced".* 80
5.3 *Example application that demonstrates the use of decision DA information. The bottom right component shows a set of decision DA extracts that are representative of the design decision of "how to find (the remote) when misplaced".* 81
5.4 *Steps involved in the feature extraction and classification process (adapted from Hall (1998)).* 84
5.5 *Lexical feature vectors for the example excerpt composed of decision DAs. Each vector is a binary variable representing whether or not each word has occurred in the excerpt.* 89
5.6 *Example decision-making discussion* 123

6.1 *Example of discourse segmentation in a produce design meeting.* 125

6.2 *Change of lexical cohesion scores over an example ICSI meeting. The red lines represent segment boundaries specified by human annotators.* 136
6.3 *Effects of transcript versions on LCSeg and the combined model when used to predict top-level and all (including subdiscourse) segment boundaries.* . 143
6.4 *Learning curve of the combined model over the increase of the training set size.* 159

7.1 *Example AMI browsers. Each browse is composed of three components: the playback facility of audio-video recordings (top), the transcription (lower left), and the extractive summary display (lower right). The two browsers differ in the summary presented in the display: the browser on the left presents the general-purpose summaries, and the one on the right presents the decision-focused summaries.* 194
7.2 *Example decision points of a product design meeting.* 195
7.3 *Task effectiveness as the average ratio of the decisions that are correctly found by the subjects. These ratios are obtained from all meetings in the series (with a total number of 30 decision points) and from the first three meetings (with 24 decision points).* 195
7.4 *Task effectiveness (number of decision hits) and perceived success (user ratings on understanding all decisions) as a function of media usage.* 196
7.5 *Number of decisions found by the subjects from all the meetings as a function of log-based task effectiveness measures.* 197
7.6 *Example decision-focused extracts of a product design meeting.* 197

8.1 *Example decision-making discussion* 212

List of Tables

2.1 *Characteristics of the two views of non-verbal communication processing.* . 23
2.2 *Features used for detecting hot spots (Wrede and Shriberg, 2003b), group-level interest (Gatica-Perez et al., 2005), and agreement or disagreement (Hillard et al., 2003).* 31
2.3 *Characteristics of the three categories of discourse segmentation approaches.* 32
2.4 *DAMSL classification scheme.* 39
2.5 *AMI DA classification scheme.* 40

3.1 *Meeting corpora.* 48
3.2 *Word error rates of the ASR system in the AMI and ICSI corpus.* 51
3.3 *Basic statistics of discourse segmentation annotations in the ICSI and the AMI corpus. ALLSEG segments refer to the combination of top-level and sub-level segments.* 54
3.4 *Basic statistics of the discourse segmentation annotations in the AMI corpora.* 55
3.5 *Pair-wise inter-coder agreement of annotations at the TOPSEG and the ALLSEG segments in two AMI meetings.* 57
3.6 *Average inter-coder agreement of annotations at the top-Level (TOPSEG) and those at both the top-level and the sub-Level segments (ALLSEG) in the ICSI and the AMI meetings.* 58
3.7 *Distribution of the major DA classes in the AMI corpus.* 58
3.8 *DA-related statistics of the discourse segmentation annotations in the AMI corpora.* 59
3.9 *Pair-wise inter-coder agreement of decision-focused extract annotations.* . . 61
3.10 *Characteristics of discourse segments that contain decision-related DAs.* . . 62

4.1 *Most frequent N-grams in the general language model and the decision-oriented language model.* 64

4.2 *Example 2x2 contingency table of word occurrences in decision DAs and non-decision ones.* 65
4.3 *Most discriminative lexical features (in bi-grams) selected by the measure of Log-likelihood (LL) statistics and DICE coefficient (DICE).* 67
4.4 *Most discriminative prosodic features selected by Log-likelihood (LL) statistics and DICE coefficient (DICE). DA is the minimal unit used for prosody analysis. PITCH denotes the normalized pitch value. $SLOPE_{Qn}$ denotes the slope of pitch in the nth quarter of a DA. Likewise, $SLOPE_{Hn}$ denotes the slope of pitch in the nth half of a DA. SD_{Qn} denotes the standard deviation of pitch in the nth quarter of a DA.* 69
4.5 *DA type distribution of the AMI decision-related DAs and general DAs. Significance codes: (***): $p < 0.001$; (**): $0.001 <= p < 0.01$; (NS): $p >= 0.05$.* 73
4.6 *Difference in the level of reflexivity and the number of addresses between decision-related DAs and general DAs. Significance codes: (***): $p < 0.001$.* 74
4.7 *DA class distribution of the AMI DAs that immediately precede and follow the decision-related DAs.* 74
4.8 *The number of subjective terms found in text and in meeting speech.* 75
4.9 *Decision discriminability of the subjective terms in meeting speech. The value of decision discriminability in each cell represents the Z-score of the subjective terms in that particular subjective category. Significance codes: (***): $p < 0.001$; (*): $0.01 <= p < 0.05$; (NS): $p >= 0.05$.* 76

5.1 *An overview of prosodic features used in this study.* 90
5.2 *DA-based features in dialogue model.* 91
5.3 *Topical features in topic model.* 92
5.4 *Subjective term features.* 93
5.5 *Feature subsets that are the most discrimintive and the least redundant, selected with Symmetrical Uncertainty and FCBF search.* 97
5.6 *Effects of different knowledge sources on the accuracy (Precision (P), Recall (R), F1) of detecting decision DAs from extractive summaries. The right three columns of both the training set and test set results are obtained using a lenient match measure, allowing a window of 20 seconds preceding and following a hypothesized decision DA for recognition. Baseline is the prosodic feature-based model.* 99

5.7 *Effects of different combinations of features on detecting decision-related discourse segments from extractive summaries.* 100
5.8 *Effects of combining lexical and other features on detecting decision DAs and decision DSs from extractive summaries.* 102
5.9 *Effects of different combinations of features on detecting decision DAs and discourse segments from entire transcripts* 103
5.10 *Effects of combining lexical and other features on detecting decision DAs and decision-related discourse segments. The first six columns are the results of operating the decision detection component on the whole recordings and the last six are on the part of recordings that have been previously selected as extractive summaries. ALL-LX1=PROS+DA+TOPIC.* 104
5.11 *Effects of different versions of DA class features on detecting decision DAs and discourse segments. The first three rows (Extract) are the results obtained on extractive summaries. The last three rows (AllTran) are the results obtained on entire transcripts.* 106
5.12 *Effects of ASR words on detecting decision DAs and discourse segments. The first three columns (AllTran) are the results obtained on entire transcripts. The last three columns (Extract) are the results obtained on extractive summaries.* 108
5.13 *Effects of subjective term features on detecting decision DAs and discourse segments. The first six columns (AllTran) are the results obtained on entire transcripts. The last six columns (Extract) are the results obtained on extractive summaries. ALL(+SUBJ)=LX1+PROS+DA+TOPIC(+SUBJ).* 109
5.14 *Effects of feature discriminability measures on average classification accuracy (F1) of decision-related discourse segment models that are trained with unigram features (LX1). Q1, Q2, Q3 and Q4 features refer to unigram features selected at different levels of lexical discriminability in descending order according to the chosen discriminability measure.* 111
5.15 *Effects of feature selection methods on the efficiency improvement of models. The last two columns show the size of the reduced feature subsets and the ratio of the subset size to the original set.* 111

5.16 *Effects of feature selection methods on classification accuracy of decision detection models in extractive summaries. The first group are models that are trained with unigram features (LX1). The second group are models that are trained with the combination of lexical, prosodic, DA-based, topic, motion and subjective term features.* 114
5.17 *Effects of ensemble models on classification accuracy of decision DA and discourse segment detection models in extractive summaries.* 115
5.18 *Decision-discriminative feature types and the systematic differences they capture.* 116
5.19 *Decision-discriminative feature types, how each type fairs in the decision detection task when used alone, and how the combined model (which integrates all the feature types) performs with each of these feature types removed. The mark in each cell indicates the level of difference between the leniently matched accuracy of this model and that of the combined model for detecting decision-related DAs from extractive summaries: + (better than the combined model); o (no difference); - (worse); – (at least 10% worse); — (the model does not work at all with accuracy under 0.1).* 117

6.1 *Performance comparison of the two probabilistic segmentation approaches.* 139
6.2 Effects of feature combinations for predicting topic boundaries from human transcripts. MC-B is the randomly generated baseline. 139
6.3 *Ranked list of feature relevance at the TOP and ALL level (in descending order).* 141
6.4 *Performance (P_k) of models trained with cue phrases from the literature (COLCUE) and cue phrases learned from statistical tests, including cue words (1gram), cue word pairs (2gram), and cue phrases composed of both words and word pairs (1+2gram). NOCUE is the model using no cue phrase features. The Topline is the human annotations on top-level segments.* 142
6.5 *Effects of feature combinations for predicting boundaries from ASR output.* . 144
6.6 *Notations used in phonetically recognized transcripts.* 148
6.7 *Example of speaker activity-augmented phonetic representation and its word-based representation.* 149
6.8 *Performance comparison of MaxEnt models trained with only conversational features (CONV) and with all available features (ALL).* 150
6.9 *Effects of individual feature classes on AMI discourse segmentation.* 151

6.10 *Effects of taking out each individual feature class from the ALL model.* . . . 151
6.11 *Effects of combining complementary features on AMI discourse segmentation.* 152
6.12 *Effects of word recognition errors on AMI discourse segmentation.* 154
6.13 *Effects of using speaker activity-enhanced phonetic transcripts on unsupervised segmentation. P_k and W_d measure the segmentation error rates.* SDis *measures the structural similarity of a hypothesized segmentation to the reference segmentation. The closer to zero the more similar to the reference segmentation.* 155
6.14 *Effects of speaker-activity models on the accuracy of off-topic, functional segment prediction. Under the K-TOPSEG condition, the total number of the ground truth segments at the top level (K_{top}) is given as a constraint to the segmenter for selecting a list of top K predictions from the hypotheses. Under the K-ALLSEG condition, the total number of segments in the two-layer ground-truth segment structure (K_{all}) is given. Under the unK condition, the total number of segments is unspecified.* 157

7.1 *Experimental design of the task-oriented evaluation: independent variables (IV).* 174
7.2 *Experimental design of the task-oriented evaluation: dependent variables (DV).* 174
7.3 *Possible misinterpretation of decisions resulted by ASR outputs.* 178
7.4 *Log file-based measures of task effectiveness.* 182
7.5 *Questionnaire-based measures of user perceived success and usability.* . . . 183
7.6 *Task effectiveness measures based on user-clicking behavior.* 184
7.7 *Quality assessment of the subjects' minutes. Results are obtained on a 7-point scale: the lower the score, the better the minute quality.* 184
7.8 *User perceived task success. Results are obtained on a 5-point scale (5 = agree strongly to 1 = disagree strongly).* 185
7.9 *Tukey HSD test results, with NS denoting "not significant". Cond 0: AE-ASR; Cond 1: AD-ASR; Cond 2: AD-REF; Cond 3: AD-REF.* 186
7.10 *Task effectiveness measures based on user behavioral cues.* 187
7.11 *ANOVA results of task effectiveness for subjects across all four conditions.* . 188
7.12 *The proportion of subjects who had low and high usage of audio-video recordings: Low=playing recordings less than 30 times; High=playing recordings greater than or equal to 30 times.* 190

8.1 *Summary of the effect of multimodal knowledge sources on decision detection and discourse segmentation.* 204

List of Abbreviations

ACT Speaker actvitiy

AD-ASR automatic decision-focused summaries displayed on ASR transcripts

AD-REF automatic decision-focused summaries displayed on manual transcripts

AE-ASR automatic extractive summaries displayed on ASR transcripts

AHMM aspect HMM

ALLSEG all-level discourse segment

AP adjacency pair

ASR automatic speech recognition

BET Browser Evaluation Test

CFS Correlation-based feature selection

CRF Conditional Random Fields

CTS conversational telephone speech

CTXT Contextual model

DA Dialogue act

DAG directed acyclic graph

DBN Dynamic Bayesian Network

DET Detection Error Trade-off

DICE DICE coefficient

DS Discourse segment

DV dependent variable

EM Expectation Maximization

F0 fundamental frequency

F1 Harmonic mean of P and R

FCBF Fast Correlation-Based Filter

FUNC functional segment

HMM Hidden Markov Model

HTD hierarchical topic detection

Hyp Average number of hypothesized segment boundaries

IBIS Issue-Based Information System

ID Industrial Designer

IG information gain

IV independent variable

LL Log Likelihood ratio

LSI latent semantic indexing

LX Lexical model

MALTUS Multidimensional Abstract Layered Tagset for Utterances

MAP maximum a posteriori

MaxEnt Maximum Entropy

MD-REF manual decision-focused summaries displayed on manual transcripts

MDD Meeting Decision Detector

ME Marketing Expert

MOT Motion model

MRDA Meeting Recorder Dialog Act

OCW Open CourseWare

P Precision

PCA principal component analysis

pLSA probabilistic latent semantic analysis

PM Product Manager

PMI Point-wise Mutual Information

PROS Prosodic model

R Recall

RSS Real Simple Syndication

RST rhetorical structure tree

SPK Speaker movements

Spurt consecutive speech with no pause longer than 0.5 seconds

su.idf Speaker-dependent inversed document frequency

SUBSEG sub-level discourse segment

TAS Twente Argument Scheme

TBET Task-based Browser Evaluation Test

TDT Topic Detection and Tracking

tf.idf Term frequency inversed document frequency

TOPSEG top-level discourse segment

TREC Text REtrieval Conference

UI User Interface Designer

VSM vector space model

WER word error rate

X2 Chi-Squared statistics

Chapter 1

Introduction

1.1 Background

This thesis, in a broad sense, tackles the problem of multimedia search and content management in audio-video recordings of human conversations. With the advances in recording and storage technology, it is becoming quite common to record everyday events. The audio-visual recordings are often stored for internal use only. In the pre-Internet era, only a small portion of these recordings, e.g., congress hearings and important press conferences, were made accessible to the public through traditional channels such as broadcast television.

In recent years, the prevalence of streaming and syndication technology, e.g., Real Simple Syndication (RSS), has enabled the Internet as a new distribution channel. Streaming has lowered the cost of distribution and, in turn, resulted in a surge of online interest in sharing videos. Moreover, since video cameras are being incorporated into increasingly diverse devices (e.g., mobile phones), non-professionals can now record their opinions easily and run media campaigns online. Everyday hundreds of millions of recordings are shared through social media sites such as YouTube[1], video blogs, and other Web 2.0 services such as Podcasts[2].

In the single month of November 2007, 95 billion videos were watched online, and over three quarters of U.S. Internet users are now using online video services, averaging 3.25 hours of video per person in November 2007 (comScore Inc., 2007). According to a February 2008 report, 52 percent of Americans who are 12 or older

[1] YouTube is a website that provides video file-streaming services. In July 2006, YouTube has reported to hit 100 million video streams served per day.

[2] Podcasts refer to the audio or video files that are syndicated for playback on computers or mp3 players such as iPod.

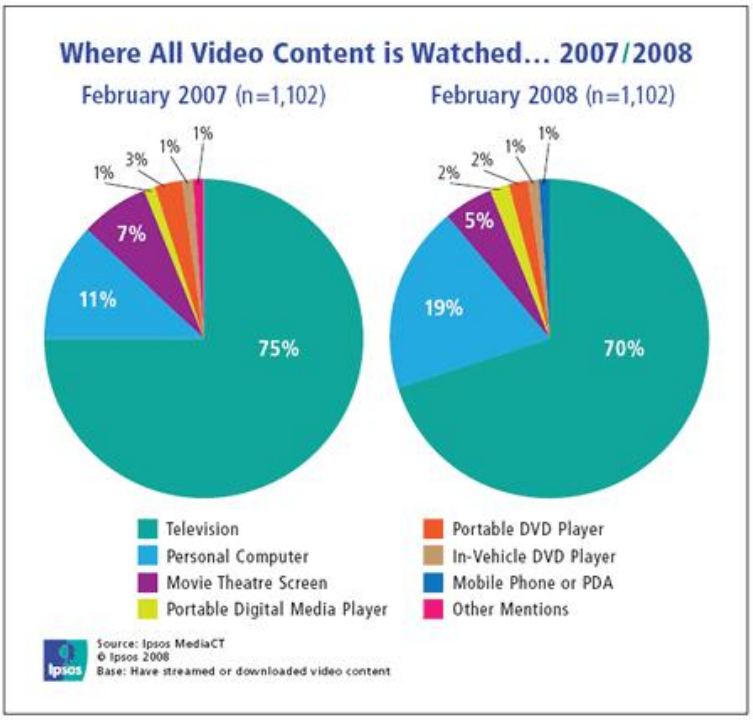

Figure 1.1: *Illustration of the increasing importance of online videos.*

have streamed or downloaded video content, and online video watching has grown to 19% of all video watched in 2008, up from from 11% in 2007. Roughly one out of every five hours of video content is now watched on a PC (as shown in Figure 1.1 (MediaCT, 2008)).

From the organizational point of view, the new technologies provide a means by which past knowledge could benefit present activities and increase productivity (Maier and Klosa, 1995). More and more academic institutions and private companies have their lectures, conferences, or even investor/analyst briefings recorded and broadcast online. For example, the MIT Open CourseWare (OCW) project shares the audio-visual content of more than 200 courses to the worldwide audience online[3]. Scientific repositories, such as SciVee[4] and ResearchTalk[5], provide open access to research talks about breakthrough ideas and findings.

Because of the emerging interest in online content, traditional media have sought to forge partnerships with the online distribution channels. For example, Fox Interactive

[3]Around 200 out of 1,800 courses have audio-video contents associated with them. Available at `http://ocw.mit.edu/OcwWeb/web/courses/av/index.htm`.

[4]`http://www.scivee.tv`

[5]`http://www.researchtalk.co.uk/rt/`

Figure 1.2: *Example audio-visual archives of various types of speech. In the clockwise order from the lower left corner are SciVee, Blinkx, YouTube, and an internal research seminar.*

Media[6], CNN and YouTube partnered to host the U.S. Democratic presidential debate in July, 2007[7]. NBC has entitled YouTube users to share its content online, as long as it is a teaser version and the users include links to the original content on its own channel[8]. CNN is also planning to provide news streaming services in Second Life to broadcast user-generated self-reporting news clips[9].

Although an astronomical number of publicly available recordings have been accumulated, to date we have not fully realized the potential of these multimedia archives. For those recordings stored for internal use, there is an "*out-of-sight, out-of-mind*" problem: With time, we are likely to encounter much more information than we can organize properly for later retrieval. Moreover, the sheer amount of data has made it impossible to find relevant pieces of information just by navigating through the content or watching to the audio-video recordings. As a result, useful information is often lost over time.

[6]Fox Interactive Media (FIM) is a leading social networking, entertainment, sports and information sites. Its network include MySpace, FOXSports.com, RottenTomatoes, etc.

[7]`http://www.youtube.com/democraticdebate`

[8]HuluDotCom. `http://ca.youtube.com/profile?user=huluDotCom`

[9]`http://www.cnn.com/2007/TECH/11/12/second.life.irpt/`

We need better multimedia search and content management mechanisms to mediate access to these archives. Existing methods have their limitations. The most popular multimedia search engines search with text-based indices, which are mined from the filenames and user-specified keywords that have been associated with the recordings. If the keywords do not reflect the content in the recordings accurately, this search method will fail. To remedy this, some multimedia search engines (e.g., Blinkx in Figure1.2) improve the method by also indexing the sound track, using the automatic speech recognition technology. In addition, other search engines improve the method by also indexing the semantic information about video content that can be automatically recognized from the videos using computer vision technology, e.g., a person standing besides a car. The index-based methods have been shown to be effective in annotating broadcast news for quick access.

However, these methods are still not sufficient to search for information in speech that is naturally argumentative, e.g., congress debates, political commentaries, and about argument development, e.g., lectures. For processing argumentative archives to answer questions such as "what are the arguments to support the government's efforts on the credit crunch crisis" or "what are the video bloggers' opinions about the new iPhone", more intelligent search (e.g., search for arguments) are needed.

Herein arises the necessity for a user-centric search facility, which allows users to refine search results and organize archives by argumentative intent for quick retrieval and easier management. The key to such a facility is an argument analysis tool that can infer speakers' argumentative intent from the recorded multimedia context.

Throughout this chapter, we will describe the problems resulting from the lack of argumentative intent understanding in this domain and why this development is interesting from the viewpoint of both application and research. Then we will provide an overview of the argumentative speech understanding research and outline the two major tasks involved, setting the scene for further discussion of the challenges and limitations facing this research.

1.2 In Search of Decisions in Meeting Speech: Application Needs

The domain we choose to examine further is multiparty meeting understanding. Meeting dialogues, which have clear argument outcomes and application needs, provide

a natural platform for this study. Repositories of audio-visual recordings of meeting dialogues constitute a valuable source of information for future training and group decision support (Romano and Nunamaker, 2001; Post et al., 2004).

The "out-of-sight, out-of-mind" problem certainly exists in this domain. While it is straightforward to record a meeting, it is more difficult to remember where the important discussions are and to find the needed information from the ever-expanding archives. Prior research has shown that standard meeting browsers, which come with simple information retrieval and playback facilities, could only help users find information relevant to ***less than 20%*** of their queries (Pallotta et al., 2007a). And this assumes that the speech has already been transcribed into text.

In responding to the failure of standard meeting browsers, several projects, to name a few, CHIL[10], NIST (Garofolo et al., 2004), IM2/M4 (Marchand-Mailet, 2003), CALO[11] and AMI (Carletta et al., 2005), have studied how to improve the browser for quicker access to the relevant discussions in meetings.

A number of these browser interfaces have been examined in Banerjee et al. (2005) for their effects on meeting information retrieval. Among those examined, the thematic (i.e., topics) and contextual (e.g., speaker roles, meeting states[12]) annotations were found to be the most effective in helping users to answer questions. However, as found in a query elicitation study (Pallotta et al., 2007a)[13], thematic and contextual queries compose only 40% of the most commonly seen user queries, whereas argumentative queries (including process and outcome) compose the majority (***60%***). In addition, recent user query analysis and organizational studies have highlighted the argumentation process and outcome as the most sought-after information among all user queries (Lisowska et al., 2004; Cremers et al., 2005; Pallotta et al., 2005) . Further user studies have suggested that decisions are the main target of user queries (Romano and Nunamaker, 2001; Post et al., 2004; Banerjee et al., 2005; Rienks et al., 2005; Pallotta et al., 2007a).

For example, knowing the decisions (i.e. argumentation outcomes) in a meeting helps users remind themselves of a meeting conclusion or audit an unattended meeting.

[10] `http://chil.server.de/servlet/is/104/`

[11] `http://www.ai.sri.com/project/CALO`

[12] Meeting states include discussion, presentation and briefing.

[13] The survey is based on the 270 queries in a simulated meeting set, IM2 (Marchand-Mailet, 2003), and the 35 queries in a survey in a natural business setting, Manager Survey set. The proportion of the queries with respect to the query classes are as follows: (i) factual level (what happens: events, timeline, actions, dynamics): 20-25%; (ii) thematic level (what is said: topics discussed and details): 11-20%; (iii) argumentative level (which/how common goals are reached): 55-63%.

Because standard meeting browsers are not sufficient for finding answers to argumentative queries, this thesis aims to develop additional plug-ins to guide the users directly to the parts of meetings that pertain to decisions. Recently, Whittaker et al. (2008) has shown that displaying key information extracted from meetings is an effective way to access meeting data. In addition, Murray (2007) has demonstrated the effectiveness of a plug-in that displays summaries, which reflect the important parts of meeting dialogues. In this thesis, we propose to study the effectiveness of another plug-in that display summaries that are related to **decisions** (as illustrated in Figure 1.3).

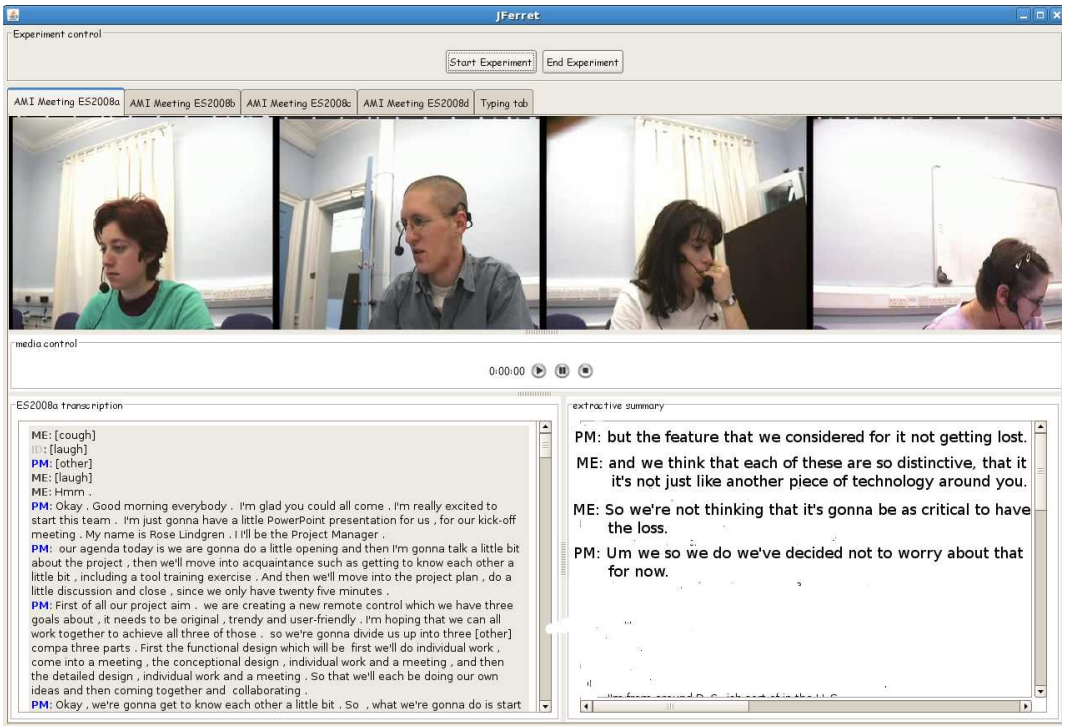

Figure 1.3: *Example meeting browser augmented with the plug-in of meeting excerpts. The bottom right plug-in displays a set of meeting sequences that are representative of group design decisions. When users click on any one of these selected sequences, the focus of the meeting transcript shown on the bottom left will be switched to where the sequence of interest is located.*

1.3 Research Questions

The development of the decision detection tool involves at least the following three categories of research questions.

- The first major category centers on **how to identify decision discussions** and

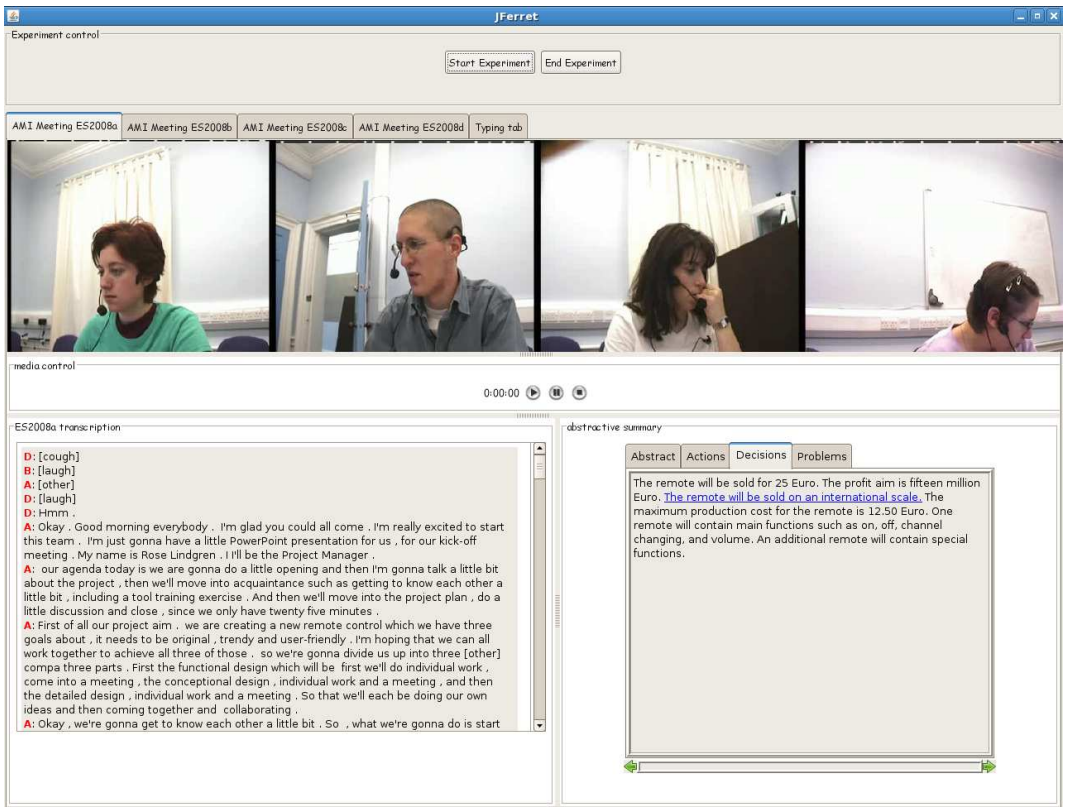

Figure 1.4: *Example meeting browser that is augmented with the plug-in that displays human-written meeting summaries. The bottom right plug-in displays the abstracts that include the decisions made in the meeting. When users click on any one of the decisions in the abstract, the focus of the meeting transcript will be switched to where the participants are discussing the selected decision.*

how to extend previous research in finding topically similar and visually similar contents to identify argumentative results.

- The second major category deals with **how to model the multimodal and multiparty cues** to identify the phenomena of interest (in this thesis, *decisions*).

- The third major category of questions addresses the issue of **how to select discriminative cues from the wide range of inputs** ("feature selection") and **how to leverage the feature structures in computational models ("structure prediction") to yield predictions** on the argumentative outcomes.

1.3.1 What techniques can be adapted to structure the multimedia archives of meeting speech?

The key challenges to developing the tool arise around the issues of how to automatically structure the multimedia archives of argumentative speech for quick retrieval. Based on previous research, three possible categories of techniques are worth investigating further: automatic indexing, automatic summarization, and emotion/private state detection.

Automatic Indexing: The first category, also the one most commonly used in the literature, is automatic indexing, which involves the generation of content-based indices using statistically derived intermediaries. Many different intermediaries have been proposed in the DARPA-sponsored Topic Detection and Tracking (TDT) workshop (TDT-Evaluation, 2002) and the follow-up TRECVID track[14]. For example, "topic models" can be derived from newswire or audio sources to index an incoming stream of broadcast news. These indices can then be used to recognize the onset of an event (e.g., Oklahoma City bombing) and to find the subsequent stories from the incoming stream (Garofolo et al., 2000; Allan, 2002). In addition, "hierarchical semantic concepts" can also be derived to index video shots of a certain location, person, thing, or event (Hauptmann, 2006).

Automatic Summarization: The second category of techniques is automatic summarization, which involves finding extract-worthy contents from the event recordings. For example, the overall statistical importance (e.g., the TF/IDF score in the vector space model (VSM)) and the semantic prominence (e.g., as a by-product of latent semantic indexing (LSI)) are often used to determine the level of extract-worthiness of the audio-video sequences.

Emotion/Private State Detection: From the previous two categories of techniques, one may suppose that previous research solely study how to organize the archives by contents. However, these content-based approach can only satisfy some of the user information needs. A third category of techniques, which is to detect emotions or speaker's private states[15]. Example end-user applications include transferring frustrated customers from a dialogue system to a human operator (Ang et al., 2002) and fine-tuning the strategies of a computer tutoring system for unsatisfied students (Forbes-Riley and Litman, 2006).

[14] Video Track in the NIST-sponsored Text REtrieval Conference (TREC).

[15] I.e., mental or emotional state which cannot be directly observed or verified (Quirk et al., 1985).

Despite the success of detecting emotions in some applications, in the context of meeting information retrieval, affective features only appears to capture shallow observations, e.g., "hot spots" wherein participants have a high level of affect in their voices (Wrede and Shriberg, 2003b,a).

Because none of these text- or speech-based techniques alone can properly describe meeting dialogues, hence in this thesis, we will complement the first two categories of techniques with the third one. While the content-based techniques respond to the factual and thematic queries, the multimodal techniques help capture some more abstract information, e.g., emotions. In particular, we will combine these techniques to solve the problem of detecting the speakers' argumentative outcomes (i.e., decisions) from speech.

1.3.2 What multimodal and multiparty cues can be used to identify decision discussions?

To integrate information from multiple communication modalities, it is important to learn "how the speakers speak and interact" in different circumstances. Previous research in spoken language understanding has shown that the differences are systematic and can be captured by modeling a wide range of multimodal cues. For example, speakers do talk in some acoustically different ways when holding different private states (e.g., frustration, uncertainty, satisfaction). The problem of how to model ***prosodic cues*** has been studied in several dialogue genres, such as goal-oriented telephone conversations, e.g., customer calls, spontaneous conversations, e.g., in Switchboard (Godfrey et al., 1992), and human-computer dialogues, e.g., in the flight-booking Communicator (Eskenazi et al., 1999) and the tutoring ITSPOKE system (Litman and Forbes-Riley, 2004). The prosodic models have lent support to the development of many aforementioned emotion detection applications (Ang et al., 2002; Batliner et al., 2003; Forbes-Riley and Litman, 2006). Ang et al. (2002) build prosodic models to detect customer frustration. Litman and Forbes-Riley (2006) train a model using prosodic and cue phrase features to detect student learning attitudes.

In the context of meetings, an even wider range of ***multimodal and multiparty cues*** are embodied in the dialogues. These cues may or may not be accompanied by words. Meeting participants often use head movements or gestures without explicitly spelling out what they mean. They also engage the whole group with long pauses and exchange glances to confirm a conclusion that has just been stated. Previous research in meet-

ing understanding has proposed many different features to model these cues to detect emotions and private states in meeting speech. For example, paralinguistic features (e.g., spectral, and disfluencies) are used in Graciarena et al. (2006) to detect deceptive speech; pragmatic features (e.g., dialogue acts and adjacency pairs) are combined with prosodic features to detect agreement and disagreement in the argumentation process (Hillard et al., 2003; Galley et al., 2004) and hot spots (Wrede and Shriberg, 2003b,a); and video features (e.g., hand and head blob position) are combined to detect group-level activities (McCowan et al., 2005).

Many of these features have also been used in other multiparty dialogue understanding applications, such as summarization (Zechner, 2002; Murray et al., 2005), topic segmentation (Galley et al., 2003), automatic detection of group-level activities (McCowan et al., 2005; Reiter and Rigoll, 2005; Zhang et al., 2005), participant roles and addressees (Banerjee and Rudnicky, 2006; Jovanovic et al., 2006), and action items (Purver et al., 2006). Therefore, in this thesis, we will examine the effect of these multimodal and multiparty interactive cues on inferring argumentative outcomes. for pattern finding.

1.3.3 How to integrate information that comes in from multiple communication modalities?

Humans understand information that comes in simultaneously from multiple communication modalities. Central to this capability is an attention mechanism that can select the most important feature subsets from the visual and auditory sensors to identify the phenomena of interest (Desimone and Duncan, 1995; Simons and Levin, 1997; Itti and Koch, 2001). Computational models can imitate the attention mechanism by applying statistical or information theoretic criterion to select relevant feature subsets. However, the feature-independent selection criteria do not consider the role of feature structure.

To remedy the lack of structural consideration, two approaches, rule-based and machine learning, have been proposed. The first approach tackles the problem by identifying interpretation rules or deriving finite-state grammars for multimodal interpretation. The identified rules or grammars are then used to enable empirical analysis (Bolt, 1980; Neal and Shapiro, 1991; Oviatt et al., 2005) or support automatic parsing of multimodal inputs (Johnston and Bangalore, 2000; Rickert et al., 2007). Although these studies are based on short dialogues and a limited set of input modes (basically, speech and gestures), it is a starting point for more advanced studies on automatic mul-

timodal information fusion. Inspired by these studies, this thesis also explores how to effectively infer feature structures from a large number of multimodal features. Due to the heterogeneity of the communication modalities in nature, this is a non-trivial task.

The second approach handles the large number of features such that models can be trained for prediction. Some of these models place probabilities over both observations and hidden structure. For example, Gatica-Perez et al. (2005) employed a layered Hidden Markov Model (HMM) to identify the group interest level manifested by interlocutors. Other models place a probability over the hidden structure given the observed feature values. For example, Hillard et al. (2003) used decision trees to classify group interest levels. Galley et al. (2004) trained Maximum Entropy (MaxEnt) models to predict whether an agreement or disagreement has taken place in an adjacency pair (AP). (Chapter 2 presents a more complete review.)

In short, these three categories of questions (with the first on "*technique adaptation for natural dialogue understanding*", the second on "*multimodal feature extraction*", and the third on "*multimodal feature selection and structure prediction*") imply a genuine need for a better multimodal information fusion framework. This framework is expected to capture the systematically different patterns that encode certain argumentative intentions (e.g., reaching decisions) such that an argumentative speech understanding component can be developed, combining information from both contents and other communication modalities.

- opening
- presentation of prototype(s)
- evaluation of prototype(s) **
  - -+ how to find when misplaced *
  - -+ preferred prototype *
  - -+ extent of achievement of targets
- target group
- costing *
- evaluation of project
- ideas for further development
- evaluation of project

Figure 1.5: *A list of major discussions (-) and sub-segments (-+) in an example meeting. The number of asterisks (*) indicates the number of decisions within each discussion segment.*

In the discussion of "how to find when misplaced" under the major discussion "evaluation of prototype(s)"
... (1) A: but um the feature that we considered for it not getting lost. ... (5) B: and we think that each of these are so distinctive, that it it's not just like another piece of technology around your house. ... (8) B: So we're not thinking that it's gonna be as critical to have the loss ... (11) A: Um we so we do we've decided not to worry about that for now.

Figure 1.6: *Example excerpt of the decision made in the discussion about "how to find (the product) when misplaced".*

1.4 Meeting Decision Detector (MDD): Task Overview

In this thesis, the use of the multimodal information fusion framework is demonstrated in the Meeting Decision Detector (MDD), a system in which meeting archives are automatically structured for quick retrieval of argumentative outcomes (i.e., decisions). In this system, the communicative patterns are detected from both the multiple modalities, the pragmatic contexts in natural dialogues, and the contents being conveyed. The features, based on both the communicative patterns and the contents, are used as indicator functions of properties of the input and the particular argumentative intent. The system models the input from various modalities, learns multimodal cues that are most discriminative of the discussion about argumentative outcomes, and structures the identified cues to predict the argumentative intent from unseen meeting sequences.

The application of MDD that we envisage in this thesis is as follows: First, a user can get an overview of what have been discussed in the missed meeting by browsing through a system-generated discourse segmentation. An example meeting discourse segmentation is demonstrated in Figure 1.5. The segmentation has two layers: top-level and sub-segments. Each segment is labeled with a topic that describes the main discussion. The more asterisk marks associated with a segment label, the more decisions have been reached in the segment. The segments that do not have any asterisk marks are those that do not contain any decisions. Take the meeting in Figure 1.5 for example. The participants have reached decisions in only two out of the eight major

segments: "evaluation of prototype(s)" and "costing". Diving into the sub-segments of the "evaluation of prototype(s)" discussion, we can then find out what exactly are the topics they have made decisions about – "how to find (the remote control the team is designing) when misplaced (in a messy living room)" and what is the "preferred prototype" the group has chosen in the end.

Next, if a user is interested to know more about a particular decision, the user can then click on the segment and be led to read through a system-generated decision excerpt. Basically, this excerpt is the highlight of the decision-making discussion – a short list of the meeting sequences that are important to understand the decision of interest to the user. As shown in Figure 1.6, those short-listed meeting sequences will be highlighted in the transcript. Supposing a usability engineer wants to find out more about the decision on the design of "how to find (the remote control) when misplaced", this user can click on the segment label to read through the highlighted points of discussion in Figure 1.6. This user will then learn that the group has decided "not to worry about designing a function to find the remote when misplaced".

To support such an application, MDD involves the following two subtasks, each of which explores one aspect of the problem:

1. automatically generating excerpts pertaining to the decisions reached in the discussion ("decision detection")

2. dividing a sequence of multiparty dialogue into segments each of which interprets a single decision ("discourse segmentation")

Figure 1.7 illustrates the flows of the two subtasks in the MDD system.

1.4.1 Decision detection: Generating decision-focused excerpt

The goal of this subtask is to locate the decision points, i.e., those parts of meeting discussions where the participants have reached some decision(s). To train a model to detect where the decision points are, the system first has to identify a set of features that can characterize decision-related discussions. This assumes that people do speak and/or act differently when they are reaching decisions. (In Chapter 3, we will examine the validity of this assumption.)

The problem can then be formulated as follows: On the one hand, there exists a distinction between decision-related sequences and non-decision sequences; On the other hand, many different types of features are extractable from the recordings, with

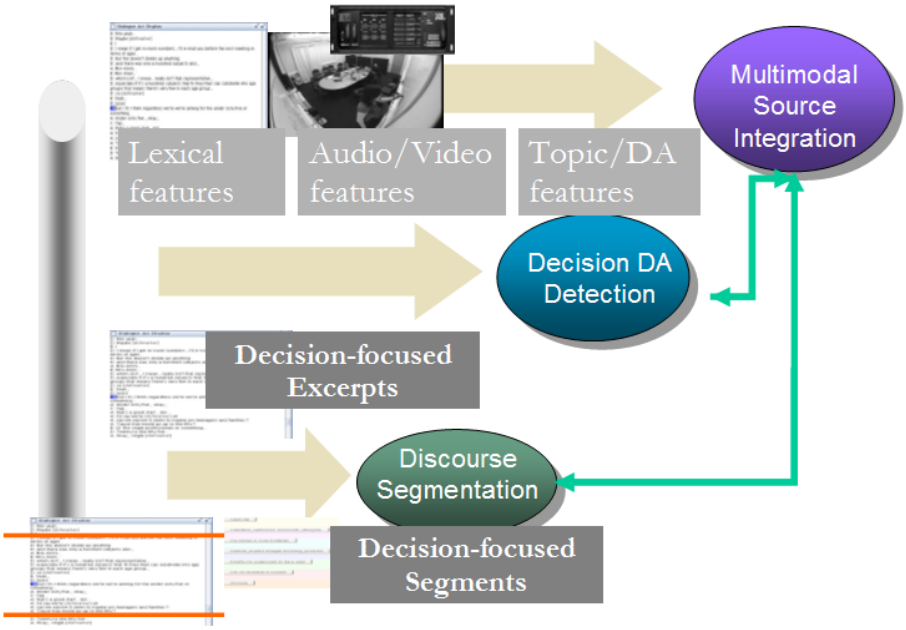

Figure 1.7: *Steps involved in the Meeting Decision Dection system.*

some representing the content aspect and the others characterizing the multimodal aspect of each meeting sequence. The key to the decision-related sequences, thus, is a set of model fitting criteria that correspond to the systematic patterns of communication modalities in the decision-related sequences. The main challenge facing automatic decision detection thus lies in the automatic identification of these decision model fitting criteria.

In this subtask, two levels of decision-related meeting sequences are detected: fine-grained decision-related dialogue acts (decision DA) and coarse-grained decision-related discourse segments (decision DS).

We assume that discourse segments would be useful for interpretation for two reasons: First, the procession of meetings often follow the agenda. Second, the topic of the discourse segments can be used as labels for interpreting decisions.

There are many different ways to display the context of a decision to the users. For example, the system can display the concatenated decision-related dialogue acts in one place as a summary and provide links back to its original content for further interpretation. Alternatively, the system can also use a fish-eye view(Furnas, 1986) to display context on demand.

Either approach can benefit from a better understanding of where discourse segments start and end. Even though the fish-eye view approach does not seem to require a pre-segmented transcript as input, discourse segmentation is still necessary to iden-

tify the labels (and the key concepts) in the shown context to facilitate user interaction. Although exactly how much context is needed for the users to interpret a decision remains as an unanswered question, it is beyond the scope of this thesis. In this study, we will assume two levels of granularity of discourse segmentation as a starting point to understand the context needed for intepreting decisions.

With the assumption made, the predictions of the decision DAs are combined to infer the decision discourse segments. The information provided by the decision DAs and that provided by the decision DSs are complementary to one another: the decision DSs offer topical information about what the interlocutors are deciding on, whereas the decision DAs provide more detailed descriptions about the decision outcome and sometimes also the supportive arguments and the level of agreement.

1.4.2 Discourse segmentation: Determining relevant contexts

As pointed out in Section 1.4.1, decision discourse segments are a natural interface for users to find decisions from lengthy meeting recordings. In fact, discourse segmentation itself is also important for quick browsing of meeting discussions. When a user wants to find information from the lengthy meeting archives, presenting discourse segmentation feeds them the most essential cues they need to grasp the overall content. Furthermore, after the user has located a topic she wants to pursue, she can use the discourse segment boundaries to locate the initial and concluding discussions of the topic. The implementation of this subtask involves grouping coherent, successive meeting sequence units into segments, each encompassing meanings beyond what is literally expressed in the individual sequence units. The main challenge facing automatic discourse segmentation thus lies in the automatic identification of these segmentation boundary model fitting criteria.

As the discourse structure considered in this work is hierarchical, we explore segmentation not just as finding major discussion boundaries and the sub-discussions (SUB), but also the off-topic discussions which largely serve the purpose of smoothing the procession of a discussion rather than contributing to the discussion [16].

[16] I collapse all the different types of off-topic segments into the functional segment (FUNC) category. Examples functional segments include opening, closing, chitchat, and discussion about agenda and equipment issues.

1.5 Goals, Future, and Guide to Remaining Chapters

In this chapter, we have provided an overview of the current practice of multimedia search and management systems, which set the scene for the introduction of the Meeting Decision Detector (MDD) system as a way to tailor the general purpose systems to what the users want to know. In particular, MDD involves two functionally correlated subtasks, described earlier and summarized here:

1. Automatic decision detection: A meeting recording will be condensed into an excerpt of meeting sequences that pertain to the decisions that have been reached in the recorded discussion.

2. Automatic discourse segmentation: A continuous stream of discourse in the meeting recording is divided into a small number of locally coherent segments. The neighbouring sequences in a decision-related segment can then be used to interpret the meaning of the decision.

This thesis addresses the problem of automatically structuring multimedia archives of natural dialogue (specifically, argumentative speech) for quick retrieval and content management. A multimodal information fusion framework is demonstrated in the MDD system. In particular, it focuses on two aspects: argumentative analysis and multimodal integration. The argumentative analysis component makes up for the lack of discourse structure modeling tools in speech. The multimodal integration component provides a foundation for capturing patterns that signal significant speaker argumentative intentions. To do this, we will use a wide range of multi-modal and multiparty cues that are extractable from the audio-video recordings and report on the quantitative impact of the different communication modalities. To incorporate the feature structure among these multimodal cues, this work will also report on the effect of feature selection and ensemble modeling on multimodal information fusion.

The problem formalization and the modeling algorithms so far correspond to an optimal setting. The experiments reported in this thesis will be run in an off-line post-meeting analysis scenario. However, this is not sufficient for our ultimate goal to run MDD online – right after a meeting or when a meeting is still in progress. Therefore, this thesis also examines techniques that provide a high degree of automation. These techniques include replacing human transcripts with the automatically recognized word- and phoneme-based transcription, and using as many automatically generated versions of features (e.g., dialogue act class) as possible.

The development of the MDD system is driven by the hypothesis that providing visual aids will assist users in reviewing the meeting decisions in a multimedia repository more effectively and efficiently. To provide insights into the scale of benefit, we will make an initial inquiry into the question: How will having a decision-focused summary assist the users to fulfill their task in the organizational context? This question will be answered by experimenting with a meeting brows er plug-in which displays decision-related summaries. User satisfaction, task effectiveness and task efficiency will be examined to determine the effect of such a plug-in.

The ultimate goal of MDD is to lend support to the development of downstream multimedia search and content management applications, widely ranging from question answering ("did the group decide about including a function to find the remote? how did they come to that decision?") to information retrieval or extraction ("find me all the decisions about budget in the past three months").

The rest of the thesis is structured as follows. Chapter 2 reviews related work of decision detection and discourse segmentation in meeting dialogues and other speech genres, such as spoken monologue or bi-party dialogues. The related work includes discourse structuring, topic detection and tracking, emotional speech detection, sentiment analysis, and other multiparty dialogue applications. Chapter 3 introduces the meeting corpus and the multi-level annotations used in this thesis. Chapter 4 presents empirical analyses of cue phrases and multimodal cues that are predictive of decision-related discussions. Chapter 5 and Chapter 6 present the experimental results in the three subtasks of MDD. Chapter 7 describes a task-based evaluation of MDD, which aims to test whether having extra decision-related information is beneficial to the end users as hypothesized. Analysis of usage log files, user-perceived success, and actual task effectiveness are summarized in this chapter. Finally, the conclusions and future work are presented in Chapter 8.

.

Chapter 2

Meeting Dialogue Understanding

2.1 Introduction

One of the goals of this thesis is to provide an account of meeting dialogue processing studies. In this chapter, we describe the need for an automatic meeting dialogue understanding system and why the development of such a system is difficult. I review the necessary conditions of conversation models that have been previously proposed and show the lack of proper models for developing automatic processing techniques in meeting dialogues. we then review the empirical studies of communication modalities used in human conversations and discuss the features that can be extracted to characterize meeting dialogues. Finally, we end this chapter with a discussion of the machine learning techniques used in meeting dialogue processing and how these techniques might lend support to the development of downstream meeting dialogue understanding applications.

2.2 Processing Meeting Dialogue: the Basics

Meetings are a critical aspect of most organizations. In meetings two or more people gather to discuss a topic, hoping to reach conclusions through the communication process. This process involves a shared goal and intensive oral arguments which provide rationales for individuals' points of view. Previous organizational studies often viewed meetings as a management tool.

The preservation of meeting information, also referred to as "organization memory", is essential to facilitate the progress of meetings and the execution of meeting decisions. Repositories of the audio-visual recordings of the meeting dialogues also

constitute a valuable source of information for future training and group decision support Romano and Nunamaker (2001); Post et al. (2004). The primary goal of past research in meeting dialogue processing is thus to develop mechanisms that are able to construct organization memory. With the recent advance of recording and storage technologies, a fast growing number of meetings are archived for later retrieval. This has led to a burgeoning interest in developing meeting browsers to help users better leverage the archived meeting recordings. The JFerret Browser (Wellner et al., 2004), for instance, contains several plug-ins which work together in order to distill the higher level information. (Tucker and Whittaker (2004) have provided an overview of meeting browsers with a variety of focused areas, widely ranging from video and artifact to discourse.)

In order to develop techniques for understanding meeting dialogues, a number of corpora have been collected (Waibel et al., 2001; Cieri et al., 2002; Garofolo et al., 2004; Janin et al., 2003; Marchand-Mailet, 2003; Carletta et al., 2006). The availability of meeting corpora has enabled researchers to begin to develop descriptive models of meeting discussions. Meeting dialogues collected in these corpora often consist of task-based conversations in which a group of people assume different roles and carry out a project together. The progress of these meetings follow certain social norms. For instance, there is typically an initial part where people introduce themselves followed by an introduction of the project proposal by the project leader. The goals and available options are determined. The meeting participants perform a range of activities, including presenting prototypes or research findings, discussing problems, making decisions and assigning action items. There will be conflicts of opinion that have to be settled by arguments.

The key issue is to provide intelligent access to the meeting information relevant to the user needs (Pallotta et al., 2007a; Whittaker et al., 2008). Previous work attempted to help users formulate their queries by providing additional meeting context cues, e.g., seating arrangement (Jaimes et al., 2004), topic segmentation, speaker role (Banerjee et al., 2005). While some focus on audio, video and meeting artifacts such as slides and shared documents, others provide additional annotations of higher level information, such as answers to a specific query. User studies showed that that the contextual cues did improve meeting information retrieval (largely fact-based) when incorporated into the meeting browser (Banerjee et al., 2005).

However, for the most common queries of meeting information – i.e., decisions (Pallotta et al., 2007a) – there is no existing technology that can identify where the

relevant discussions are. In the following sections, we will review prior work in related areas which suggest the types of features and approaches that are useful for meeting dialogue understanding, with an emphasis on those that may be relevant to decision detection.

2.3 Decision Making Process Modeling

Decision-making process is the deliberation process leading to the choice of a course of actions among several alternatives. Discourse surrounding the decision-making process is communication that goes back and forth (from the Latin, discursus, "running to and from"), such as debate or argument. Such discourse prevails in our daily communication across widely ranging communication mediums in written and oral communication. Regardless of the communication medium in use, our cognitive system can process decisions effectively.

How do humans choose between alternatives has been the subject of active research from many perspectives. In practice, the findings can be used to purport the development of many knowledge-based expert systems. For example, medical diagnosis systems help select the right treatment from a number of alternatives; computer chess programs are designed to select the best move from all the possible course of actions.

Cohen (1993) summarized three paradigms that have been proposed to perform decision analysis: the formal-normative paradigm, the rationalist paradigm, and the naturalist paradigm. All of the three aim to describe how a confused decision maker, who wishes to make a reasonable and responsible choice among alternatives, does his job. In this subsection, we will review all the three paradigms, notwithstanding only the naturalist paradigm handles the problem of how this could be done within the context of a group.

Formal-normative paradigm: The mainstream formal-normative paradigm is concerned with how to leverage various functions (e.g., utility functions, loss functions, risk functions) to estimate the return (i.e., expected utility) of each possible outcome. Psychologists have performed behavioral experiments to understand how subjects make decisions under uncertainty (Mosteller and Nogee, 1951; Davidson et al., 1957; Peterson and Beach, 1967; Tversky and Kahneman, 1981). In the experiments, humans are found to be imperfect statisticians, whose decisions often deviate from the predictions of the formal-normative model. To explain why decision making behav-

iors do not always match the maximization of potential returns, numerous variations of selection strategies (e.g., conjunction, disjunction) have been proposed to realize how decisions would be reached under different requirements, e.g., maximum expected utility, maximum subjective expected utility (de Finetti, 1964; Savage, 1972), dominance (achieving least as good results as any other strategies) (c.f. the survey in (Levy, 1992)).

A few researchers have attempted to extend this paradigm to model group decision making process (Bodily, 1979; Baucells and Sarin, 2003). At the group level, multiple criteria are weighted with respect to individual utilities such that the weighted sum can be used to infer group decisions.

Rationalist paradigm: While the formal-normative paradigm focuses on the selection of best outcomes, the rationalist paradigm emphasizes more on the actual psychological steps in the cognitive process of decision making. Following rationalist philosophers' footsteps, this paradigm is rooted in the assumption that decisions are made out of rationality and therefore can be broken down into short, logically self-evident steps.

Researchers in decision analysis have developed a varity of graphical representations to model the psychological steps, e.g., multi-attribute objectives, inference diagrams, decision trees and Bayesian decision rules (Keeney and Raiffa, 1976; Howard and Matheson, 1984; Clemen, 1997; Jensen, 2001). These representations have enabled the modeling of a multi-attribute, group decision process. For the part about group decision making, theoretical models, e.g., those based on **social choice theory**, have been developed to characterize how a collective decision would be reached out of individual preferences. For the part about multiple attributes, **fuzzy sets** which considers preference orderings and relations have been applied to aggregate multiple utility functions in a multipurpose decision making model (Chiclana et al., 2002). Linear programming algorithms have been leveraged to find the optimal group decisions (Lewis and Butler, 2007).

As these decision theories still lack explanatory power in response times and outcome selection in the face of "new" alternatives, more recently researchers have proposed stochastic models to account for the context effects, e.g., similarity effect and compromise effect, of newly available options. Busemeyer and Diederich (2002) have surveyed the theory of **decision field**, in which decision making is modelled as a sequential sampling process so that the valance of each option can be integrated over time to generate output preference states. The accumulation of preference predicted by this

theory have been confirmed by the studies of neural correlates as closely related to the neural firing rates (Smith and Ratcliff, 2004).

Naturalist paradigm: However, arguments in real-world decision making processes do not always follow the normative model and the relatively ad hoc cognitive procedure. A new trend of "practical decision-making" research has thus emerged: Rather than testing how well a model fits behavioral decision scenarios in a lab setting, researchers have performed field studies to understand how professionals make decisions in a naturalistic setting. Specifically, naturalistic researchers have paid attention to how the knowledge structure are created and adapted in a real-world decision making process (Beach and Mitchell, 1978; Orasanu and Connoly, 1993; Cohen, 1993; Rasmussen, 1986). For example, Rasmussen (1986) studied the cognitive control of human tasks in a power plant and how accidents were resulted. The observed cognitive models are clssified into three levels: skill-based (i.e. subconscious movement patterns), rule-based (i.e. subroutines that prescribes actions for a situation with conscious attention), and knowledge-based control (i.e. conscious decision making).

The benefit of the naturalist paradigm remains when the level of analysis is extended to groups. The point of departure for the group-level and individual analysis lies on the goal, with the group decision-making efforts geared toward long term goals rather than immediate outcomes (Donaldson and Lorsch, 1983). In addition to the previously observed cognitive models, many more models have been proposed to handle the newly added aspsct of group evaluation. For example, researchers have developed a story-based model to describe how jury members evaluate court evidences as a group. In particular, three types of knowledge are coordinated in group evaluation: the knowledge of current situations, knowledge of similar situations, and expectations of what are needed for completing the story.

However, with the advance of naturalist studies, it has become evident that real-world decision-making processes are often too complex and uncertain to be completely covered in any of the three paradigms alone. This is also where linguistic approaches come into the scene, deriving complementary cues from conversations by taking in one piece of information, e.g., warrant, backing, evidence, at a time to compose the decision-making process. Linguists have done their best to identify linguistic quantifiers that can be used to identify decision-critical information, e.g., majority expressions (Kacprzyk, 1986; Herrera et al., 1995; Xu, 2008), from the data directly. These quantifiers range widely from simple majority cues, e.g., "most", "much more than 50%", to complex syntactic and semantic rules that are designed to determine the de-

gree of dominance.

	Language-as-product	Language-as-action
Method	info processing with generative grammar; mechanistic; decontextualized	rational analytic; non-mechanistic; developed in natural contexts
Goal	understand how humans comprehend and produce language by processing multi-level representations	understand the processes of forming joint intention (and joint action)
Development platform	Monologue	Dialogue

Table 2.1: *Characteristics of the two views of non-verbal communication processing.*

One drawback of the current linguistic approach is its sheer focus on the verbal part of the decision-making conversation. Llimited research has been done to find non-lexical quantifiers. Yet in the group context, speakers use communication mediums more than words to express meanings. In turn, the non-verbal mediums have driven group communication difficult to measure. In fact, previous research to the processing of non-verbal communication has developed quite some interesting methods. As pointed out by Clark (1996), these methods follow a historical dichotomy, adopting either the view of "language-as-product" or "language-as-action". (Table 2.1 presents the characteristics of the two views.) In short, while the former focuses more on individual cognitive processes in producing and understanding language, the latter streses the social-cognitive factors in speaker communication.

The current lingistic approach to decision-making conversation processing has mainly taken the "language-as-product" view, hence inheriting the mechanistic tradition of language processing developed almost entirely on monologue. This is also to assume that the results on monologue understanding can be extended to understand dialogue. Even the psycholinguistic studies on non-verbal communication (e.g., gesture-speech synchrony (McNeill, 2000), eye-movements around a visual scene (Henderson, 2004), prosody (Beattie and Shovelton, 1999)) also assume similar positions to understand speech planning, i.e., how the core content of a surface utterance is self-organized and discourse is extended at each moment of speaking. Pickering and Garrod (2004) have pointed out the necessity of supporting an interactive alignment mechanism that can incorporate contexts into the multi-level representations.

In contrast, cognitive and social psychologists have explored the forming of joint intentions and actions (including the use of non-verbal cues) in different contexts (e.g., exaggeration). (Please refer to Clark (1996) and Fussell and Kreuz (1999) for an overview of the methodology and case studies in this line of research.)

New trends have emerged to investigate into the possibility of merging the two views of language processing. Several case studies have been exemplified in Trueswell and Tanenhaus (2004). In this thesis, we will follow the direction to find the coupling of verbal language (not just quantifiers) and non-verbal cues that can be used to indicate decision-critical information.

2.4 Conversation Modeling

To further suggest bottom-up wayss to characterize group decision-making process, this study searches for a theory of content representation which can abstract a discussion into its inherent conceptual organization while also preserving the relations between the discourse units.

Traditionally, theories of content representation were developed by systematically analyzing the text structures larger than the sentence. For example, Harris (1963) developed a discourse analysis method to seek connected discourse units which contain some global structure characterizing the whole discourse or large portions of it. Kaplan (1972) described a rhetoric model which constructs representation of coherent units. However, most of this research focuses on characterizing what makes text coherent at the level of syntactic structure. It is also questionable how well the syntactic structures can be parsed from conversational speech, which is verbose and structured differently from text by nature.

2.4.1 Modeling what the speakers say and mean

To model the organization and development of what the speakers say, we review a variety of models of discourse segmentation that accounted for the semantic or logical relations between discourse units. In the following subsections, we follow Kehler (2002) to categorize these models into two categories: **informational coherence** and **intentional coherence**.

2.4.1.1 Informational coherence

In the first category, discourse segments are found by grouping the successive conversational units that are similar in their thematic focus (i.e., the main topic in each of the utterances). The thematic focus can be determined by looking for conjunctive relations, which usually appear in the form of an initial theme unit followed by another unit that refers to this theme unit, e.g., the-But-relation, the-Then-relation. Halliday and Hasan (1976) analyzed such lexical coherence and described a taxonomy of the conjunctive relations.[1]

Some studies referred to each unit's role or the type of content in the discourse as rhetorical predicates (Grimes, 1975; McKeown, 1985)[2] and then organized rhetorically connected units into coherence discourses. Mann and Thompson (1988) further defined a hierarchical taxonomy of 23 finer-level rhetorical relations.[3] Relations recognized with this taxonomy are then fitted onto the higher-level structures (i.e., schema applications) that are derived from lower-level structures (i.e., schema), thus enabling the formation of a rhetorical structure tree (RST). RST theory was expected to allow analyses of the majority of English discourses. (A more detailed review of the above approaches can be found in Knott (1996).)

However, it is uncertain whether these models previously proposed for analyzing written communication can be applied to analyze oral communication.

2.4.1.2 Intentional coherence

In the second category, discourse segments are found by grouping units with connected intentions. Discourse units are inter-connected by coherence relations, and the relations hold between these units to form a larger discourse segment. From the 80's, computational linguists have begun to compute discourse segments by developing recursive algorithms to find unit structures that are intentionally coherent. A variety of coherence relations[4] were proposed to construct a local structure of discourse units

[1]The proposed conjunctive relations include the additive, adversative, causal, and temporal relation.

[2]Example predicates include identification, evidence, amplification, inference, attributive, cause-effect, analogy, explanation, constituency, renaming, generalization, particular illustration.

[3]Example top-level relations include subject matter and presentational. The subject matter relation is subdivided into elaboration, circumstance, solution, cause cluster, condition, otherwise, interpretation, evaluation, restaurant, summary, sequence, contrast. The presentation relation is subdivided into motivation, antithesis, background, enablement, evidence, justify, concession.

[4]E.g., Occasion (a change in unit U_i causing or enabling U_j), evaluation, background, explanation, parallel (expanding from specific to specific positively), contrast (specific to specific negatively), generalization (specific to general), exemplification (general to specific), elaboration, violated expectation.

(Hobbs, 1979, 1985; Polanyi, 1988).

Other prior work (Grosz and Sidner, 1986; Lochbaum, 1991) applied recursive planning algorithms to find larger sizes of discourse segments, e.g., collaborative plans, in which units are rhetorically connected with dominance[5] or satisfaction-precedence relations[6].

Discourse segments found in these two categories – information cohesiveness and intentional coherence – are often posited as isomorphic (Grosz and Sidner, 1986). Cognitive studies showed that this is consistent with human cognitive process: Human cognitive systems constantly monitor informational cohesiveness and intentional coherence in discourse to group successive units into discourse segments. Depending on the medium and the application, human cognitive systems may choose to attend to one or both. Morris and Hirst (1991) provided empirical evidence on how the successive units that are intentionally coherent are mostly cohesive.[7]

2.4.2 Speech acts and collaborative plans

In addition to modeling what the speakers say, psycholinguistic and theoretical linguistic research in human dialogue understanding has also modeled other pragmatic contexts that indicate what the speakers mean, that is, the underlying meaning and implicature of the successive units in discourse structure.

A number of models, for instance, those based on individual speech acts (Searle, 1969; Grice, 1975; Grosz and Sidner, 1986), turn-taking control rules (Schegloff, 1968; Duncan, 1972)[8], and on collaborative plans (Searle., 1990; Lochbaum, 1991; Grosz and Kraus, 1993), have been proposed to model the pragmatic contexts of discourse units in dialogues, such as telephone conversations and human-computer dialogues.

However, these individual or group act-based models commonly used transcripts as the proxy of speech for analysis, and spontaneous face-to-face dialogues in meetings violate many assumptions made in these models. While processing transcripts helps capture what the meeting participants said in the meetings, processing only the tran-

[5]The dominance relation connects a discourse unit with its subordinating units whose satisfaction are intended to provide part of the satisfaction.

[6]The satisfaction-precedence relation exists if two units are dominated by another unit and the intention of one of the units has to be satisfied before the remaining unit.

[7]Morris and Hirst (1991) examined 183 sentences from general-interest magazines such as Reader's Digest. Although informational cohesiveness does not translate into intentional coherence necessarily, most of the segments found by information cohesiveness and intentional coherence do correspond to one another.

[8]Schegloff (1968) presented rules for ensuring only one speaker at a time; Duncan (1972) presented rules for preventing speakers to interrupting others.

scripts for analysis will result in missing information that is communicated through other modalities, e.g., how the participants expressed themselves and how they interacted with each other. Because we also want to capture the information conveyed through these other communication modalities in meetings, in this section we also review prior work that identified multimodal communicative patterns from speech.

2.4.3 Modeling argument intent and structure

The work on discourse segmentation forms the basis of further discussion on argument understanding. In this thesis, we focus on (1) the necessary and sufficient conditions of how an argument starts and ends, and (2) the communication modalities interlocutors use to express propositions. However, given the number of potential features available, this thesis will only attend to those that can be captured or inferred from the audio or video recordings. Before continuing it is important to clarify what is the "argument" we are trying to locate in dialogues.

Arguments have traditionally been closely associated with reasoning and language. The point of departure in the studies of argument is what the term "argument" really means. For philosophers, argument was first seen as grounding claims of the logic in the epistemologic system. Then, with the advent of Pragmatism theory, argument was used more to account for the credible arguments which establish a convincing conclusion about various issues. The term argument is also used by speech and communication scholars, such as Willard (1988) and O'Keefe (2002). Willard (1988) defined argument as "a form of interaction in which two or more people maintain what they construe to be incompatible positions".

The recovery of meeting arguments involves determining both the main topic of the arguers' incompatible positions and the outcome. As the focus of this thesis is on the argument outcome, in this subsection we also discuss some of the models in previous research that focus on recovering argument structures.

2.4.3.1 The Toulmin model

Toulmin (1958) presented a schematic representation of the procedural form of argumentation, in which positive arguments and acceptability of certain statements are considered. In this model, an argument is regarded as a sequence of interlinked claims or reasons which together establish the position for which someone is arguing. It consists of the following six building blocks.

1. Claim: the position (claim) being argued for; the conclusion of the argument.
2. Grounds: supporting evidence that bolsters the claim.
3. Warrant: the principle, provision or chain of reasoning that connects the grounds/reason to the claim.
4. Backing: support, justification, reasons to back up the warrant.
5. Rebuttal/Reservation: exceptions to the claim; description and rebuttal using counter-examples and counter-arguments.
6. Qualification: specification of limits to claim, warrant and backing. The degree of conditionality asserted.

Compared to the definition of argument given in Willard (1988), the definition used in the Toulmin model is different in the way that it only concerns positive arguments. Although it is not suitable for modeling the whole meeting discourse, whose argument processes can yield both positive and negative outcomes, it suits the development of systems that detect only those that resulted in positive outcomes.

2.4.3.2 The Issue-Based Information System (IBIS) model

Kunz and Ritte (1970) presented an IBIS model that could structure the dynamics of a problem discussion (i.e., topic). The IBIS model is an information representation system that provides a hierarchically linked web structure to manage information around issues. Specifically, it can capture participants' positions on issues and their alternatives. During an argument process, participants can construct arguments to defend their positions. The possible actions arguers can take are shown in Figure 2.1. This model has been used for task-oriented dialogues among participants who each have their own area of expertise. The definition used in this model is broader than the Toulmin model. However, its application is still limited because its topic labels have to be predefined in an ontology.

2.4.4 Features Characterizing Meeting Conversations

Prior empirical studies analyzed the multimodal contexts near segment boundaries. These studies have identified systematic differences in a variety of verbal features, e.g.,

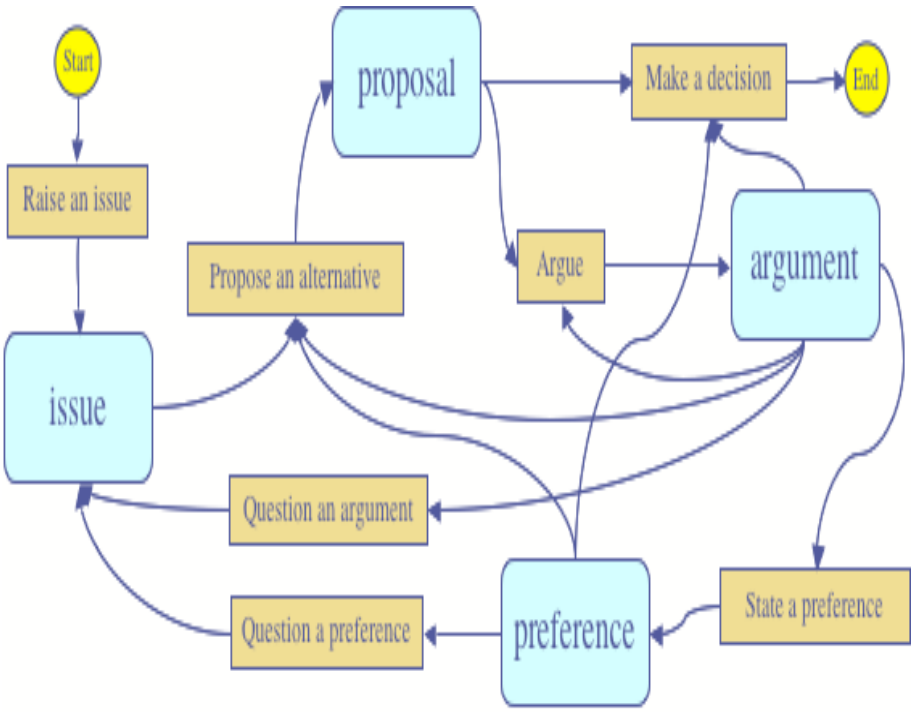

Figure 2.1: *Possible actions in the IBIS model. (Adapted from Kunz and Ritte (1970).)*

discourse connectives (Litman and Passoneau, 1995; Beeferman et al., 1999; Hutchinson, 2004), and non-verbal features, e.g., turn-taking cues (Sacks et al., 1974; Levinson, 1983), intonation cues (pitch, pause, duration, contour) (Hirschberg and Litman, 1987; Grosz and Hirschberg, 1992), intonational patterns and speech rates (Shriberg et al., 2000), hand gestures (Grice, 1969), eye gaze, and head nods (Cassell et al., 2001).

In addition, past research has also worked on identifying "speaker intention-revealing" cues in the different modalities. For example, Duncan and Niederehe (1974) studied cues that can signal the intention of participants to keep or pass the floor, or to interrupt. Other nonverbal interaction cues, such as gaze, facial expressions, posture, head movements, gestures, have been studied for their contribution to conversation flow control (Argyle et al., 1973). Dialogue systems often use these cues extracted from the preceding dialogue acts to recognize the functional role of the next conversational unit (Nagata and Morimoto, 1994; Reithinger and Klesen, 1997; Wright, 1998; Hastie et al., 2002).

However, relatively few computational modeling approaches have addressed the issue of automatically deriving discourse structure from pragmatic contexts. The em-

pirical studies focused more on the coverage of linguistic phenomenon rather than the applicability of developing computational models from the proposed features. The lack of proper computational models of discourse structure has pointed out an important direction for this thesis. Also, the proposed models of argument structure are designed to understand the user intention of the immediately preceding user moves in a task-oriented environment; The applicability of these models to free-form conversations has thus been questioned.

To integrate information from multiple communication modalities, previous research in spoken language understanding studied how to automatically derive multimodal cues that indicate "how the speakers speak and interact". For example, ***prosodic cues*** can be automatically extracted through analyzing the pitch contour and energy level over speech signals and the duration over syllables. The identified prosodic cues have lent support to the detection of various conversation phenomenon, e.g., speakers' private states (e.g., frustration) uncertainty, satisfaction) (Ang et al., 2002; Batliner et al., 2003; Forbes-Riley and Litman, 2006). Its applications range widely from detecting frustrated customers in service calls (Ang et al., 2002) to identifying student learning attitudes during tutoring sessions (Litman and Forbes-Riley, 2006).

In the context of meetings, ***multimodal and multiparty cues*** are embodied in the dialogues, accompanied by words. Previous work has studied how to automatically extract the indicative cues, e.g., head movements, gestures, long pauses, and glance exchanges. Table 2.2 summarizes the features that have been used to detect various high-level conversational phenomenon, e.g., hot spot (Wrede and Shriberg, 2003b), group interest level (Gatica-Perez et al., 2005), agreement or disagreement (Hillard et al., 2003). In Section 2.6, we review prior work that has leveraged the multimodal and multiparty cues to identify conversational phenomenon in meetings.

2.5 Toward Automatic Derivation of Discourse Structure

As previous research focused more on elucidating linguistic phenomenon, it did not necessarily consider automation. Among those that have considered automatic derivation of discourse structure, many rendered formal methods. For example, Kehler (2002) and Wolf and Gibson (2004) derived a tree structure of discourse segments by

	Hot spots	Group interest level	Agreement or disagreement
Lexical	Utterance length Perplexity (4-grams)	N/A	Spurt length[9] Content words Perplexity (4-grams) The class of first and last word Agreement markers Subjective words General cue phrases Class-indicative keywords The class of the first two words
Auditory	F0 Energy	SRP-PHAT Energy Pitch Speaking rate	F0 Speech rate Overlap Gap (silence) Duration Average, maximum and initial pauses
Video	N/A	Global person motion Eccentricity/orientation of hand blobs Head orientation	N/A
Structural	N/A	N/A	The speaker of the previous and the next spurt Number of speakers Number of spurts in-between an adjacency pair (AP)[10]
Pragmatic	Dialogue acts Speaker type Meeting type	N/A	Previous tag dependency Reflexivity Transitivity

Table 2.2: *Features used for detecting hot spots (Wrede and Shriberg, 2003b), group-level interest (Gatica-Perez et al., 2005), and agreement or disagreement (Hillard et al., 2003).*

concatenating successive units that contain coherence relations[11]. However, the formal methods are often brittle, as they render inferences across propositional contents in the

[11]Two adjacent units are considered at a time; If there exists a coherence relation (e.g., cause-effect, violated expectation, condition, similarity, contrast, elaboration, attribution, temporal sequence) between the situations described by the two units, then the two units can be concatenated as a discourse segment.

discourse and require a full-fledged knowledge base in which all the possible concepts and the relations between concepts are defined.

To remedy the lack of automatic models, statistical approaches have been proposed to find discourse structures (usually a simple flat one) in dialogues. It usually involves subdividing a dialogue into a number of smaller segments. The proposed approaches can be grouped into three main categories: **semantic clustering**, **sequence decoding**, and **feature-based boundary classification**. A simple comparison of the three categories of models is presented in Table 2.3. In our experiments in Chapter 6, we will report the use of these different approaches on the task of meeting discourse segmentation.

	Semantic Clustering	Sequence Coding	Feature-based Classification
Convergence	n/a	+	++
Computation Cost	low	very high	high
Training set (Prior knowledge)	no training	larger	smaller
Parameter estimation	n/a	complex	less complex
Feedback (fea. (contribtion)	explicit	implicit	implicit
Efficiency	Fastest	Slow	Fast
Accuracy	+		++

Table 2.3: *Characteristics of the three categories of discourse segmentation approaches.*

2.5.1 Semantic Clustering

In the first category, discourse segmentation is viewed as a time series analysis problem amenable to signal processing. This task thus involves finding important patterns from transcripts and detecting where the patterns have salient changes.

In practice, the patterns are usually determined by tracing the informational coherence over successive discourse units. The cutting threshold of cohesiveness can be found either by heuristic rules or by an automatic mechanism. For example, the heuristics observed in the empirical study conducted by Morris and Hirst (1991) yielded an

algorithm which predicted segment boundaries in text. The heuristics were used to determine what words to be included in the lexical chains in the first place, what thesaurus categories the words belong to, and whether the words in the lexical chains belong to same or related categories or exhibit transitive relations. Reynar (1998) proposed another algorithm to predict segment boundaries in news stories, using automatically derived lexical cohesiveness scores.

In following this attempt, Hearst (1997) defined patterns as lexical chains across the units and determined coherence by analyzing the number of overlapping chains. She proposed TextTiling, an unsupervised approach which seeks to place discourse segment boundaries at points where the lexical patterns change noticeably. This approach has been extended by Stokes et al. (2004) and Matveeva and Levow (2007) to hypothesize segments in broadcast news and multiple documents.

Important patterns can also be determined through dimension reduction algorithms, such as multinomial principal component analysis (PCA), which aims to find coherent subsets that are relatively independent of each other. A wide range of clustering algorithms such as K-means, latent semantic analysis (LSA) (Deerwester et al., 1990) and probabilistic latent semantic analysis (pLSA) (Hofmann, 1999) have been applied to group semantically similar units.

With the semantic groups automatically identified, the task of discourse segmentation can be recast as an optimization task, employing graph-cutting algorithms to identify the best partition of units, for instance, those that minimize the level of inter-partition similarity without compromising the intra-partition similarity. Several algorithms have been proposed to determine an optimal segmentation, including dynamic programming (Ponte and Croft, 1997; Utiyama and Isahara, 2001), agglomerative clustering (Yaari, 1997), and latent semantic analysis (Choi et al., 2001). The main difficulty is to automatically determine the number of segments for an optimal segmentation.

These graph-based algorithms can be successfully extended to find "hierarchical" discourse segmentation. For instance, previous work has presented a number of algorithms. For example, directed acyclic graph (DAG)-based segmentation (Trieschnigg and Kraaij, 2005) works well in hierarchical topic detection (HTD), the goal of which is to organize an unstructured discourse (e.g., news collection) into its topic structure. (For more related work, please refer to the report of the hierarchical topic detection task reported in TDT (Trieschnigg and Kraaij, 2004).

2.5.2 Sequence decoding

In addition to the semantic similarity-based algorithms, the task of optimizing discourse segmentation can be achieved via sequence labeling. Here, we model each discourse unit as being generated from a particular distribution over topics, which is assumed with a Markov structure, such as Hidden Markov models (HMM).

Many algorithms have been proposed to find a topic label sequence that best encodes information in the consecutive observations (i.e., discourse units). The goal of this generative approach is to find the maximum joint probability of the paired observation and topic label sequences by enumerating a finite number of sequences.

The enumeration process usually consists of the following steps: First, clustering is applied to construct k (smoothed) N-gram models from a large corpus (e.g., the Wall Street Journal, the CNN news transcripts), with each model $T^{(j)}$, $1 \leq j \leq k$ representing one specific topic. Next, each discourse unit is considered as a collection of mutually independent words (with its length as L) that are generated from a topic model z. More formally, each observation is represented as a word vector in a fixed analysis window: $o_t = w_{t,1}, w_{t,2}, w_{t,3}, ...w_{t,L}$, and each state of the HMM is represented by a topic variable (z_t), which is sampled from a number of topic variables. In the simplest case, the set of topic variables may only contain two states as in the binary classification framework: discourse segment boundary (YES) or not (NO). The emission probability, i.e., the likelihood of each of the consecutive utterances being generated by this topic model, is calculated with respect to each of the k topic models:

$$P(o_t|z) = \prod_{n=1}^{L} P(w_i|z) \tag{2.1}$$

The topic that yields the highest probability is then selected as the topic of the unit. Transition probabilities among the topics and the self-loop probability are also calculated. Finally, topic breaks occur at the points where the value of the topic variables for the next utterance changes ($z_t \neq z_{t+1}$).

Based on these calculations, the task of segmenting discourse is thus cast as that of locating the change of topics between the current unit and its successor (van Mulbregt et al., 1999; Blei and Moreno, 2001). Although these researchers found that the HMM framework performs well on segmenting broadcast news, there are limitations inherent in the framework: First, the framework imposes some over-simplified Naïve Bayes assumptions on the interactions between hidden states and observations: the set of words in each observation are sampled i.i.d. given the current topic, and each state depends

on only the previous topic. Second, while these HMM variants are typically trained to maximize the joint probability of the observations and the states, the end task is often evaluated with the posterior probability of the state sequence given the observations. Finally, long-range dependencies are difficult to model in this framework due to the intractable problem incurred by the inference step and the number of parameters in the model that is too large to be learned from data.

HMM variants have been proposed to cope with these limitations, such as aspect HMMs (AHMMs) (Hofmann, 1999), integrated HMMs, Conditional Random Fields (CRF) (Lafferty et al., 2001). Some have applied the variants to solve problems that require a segmentation component, e.g., finding topical sentiments from a large amount of webpages (Hurst and Nigam, 2004), determining utterance boundary sites in broadcast news (Liu et al., 2004) and topic segment boundary sites in written news and radio programs (Blei and Moreno, 2001).

2.5.3 Feature-based classification

As opposed to the previous two categories of approaches that consider lexical cohesive discourse units, a supervised, discriminative modeling approach in the third category is applied to identify the disruption points of lexical cohesiveness and intentional coherence, that is, where the segment boundaries are. This approach trains a classification model to distinguish discourse segment boundaries from non-boundaries. The task of discourse structure is decomposed into a series of binary classification problems, determining whether each of the discourse units in the discourse is at the boundary of any discourse segment.

Because the classification approach does not impose a rigid constraint on the type of features that can fit into the model, is has been used to combine not just the lexical features, which are representative of content, but also a large number of multimodal features, which are important to the recognition of speaker intention. Liu et al. (2004) have provided a detailed account of the effect of some possible feature combinations, e.g., those incorporating prosodic and lexical features, on many different spoken language understanding tasks, such as finding utterance breaks and detecting pauses and disfluencies.

The discriminative classification approach can be formalized as follows. For each discourse unit (i.e., the potential candidate of a segment end), a number of verbal and non-verbal features are extracted from the unit's context X. Given X, a pre-trained

model $Q(y|X)$ is used to classify which boundary class y, where $y \in \{YES, NO\}$, this unit belongs to.

A variety of machine learning algorithms Q have been proposed to train the classifier from training data in research in spoken narratives and short dialogues. The same approach can be applied to segment meeting monologues. For instance, Q can be a decision tree, i.e., a set of decisions rules. Previously, Grosz and Hirschberg (1992) trained a regression tree CART from acoustic features (e.g., pause, duration and contour) to place segment boundaries in spoken dialogues. Litman and Passoneau (1995) used another decision tree implementation, C4.5, with a combination of prosodic, cue phrase, noun phrase, and cue-prosody features to perform segmentation on spoken narratives.

Q can also be an exponential model, i.e., a decision function parameterized by feature weights. For example, Beeferman et al. (1999) and Christensen et al. (2005) trained exponential models with both verbal features (e.g., the occurrence counts of topical words and discourse connectives in a neighboring window) and non-verbal features (e.g., pause duration) to segment broadcast news.

Beeferman et al. (1999) constructed an exponential model to select useful features from a set of cue phrase features and topicality features, which were trained from both long-range and short-range language models). Christensen et al. (2005) improved this approach by constructing a maximum entropy model and considered pragmatic contexts, e.g., by including prosodic features. (For the complete set of features used, please refer to Chapter 6.)

2.6 Meeting Dialogue Processing

Despite the lack of theoretical models that can operate on free-form conversations in meetings and that consider the pragmatic contexts (which are indicated by the non-verbal features), this thesis still believes that some of the models discussed here can be extended to understand conversational phenomenon in meetings. In addition, this thesis hypothesizes that there do exist systematic patterns of communication modalities in the "structure-signalling" and "speaker intention-revealing" discourse units. Once we have the key communication modalities identified, discourse models can then be automatically generated by learning the fitting criteria from data. In the following sections, we review related research in the context of meeting dialogue processing.

2.6.1 Hierarchical segmentation

As discussed in Section 2.4, current models of discourse structure are constrained to interpret the immediate context that can be inferred from the words spoken. However, the meeting dialogues are often lengthy, and speakers use many multiparty and multimodal cues to imply meanings beyond words. Therefore, the interpretation of these lengthy dialogues requires a longer memory span that takes into account information from all available communication modalities.

The high requirements have limited work on recovering meeting discourse structure to recovering discourse segmentation. To tackle the problem of discourse segmentation recovery, the three categories of approaches discussed in Section 2.5 have all been applied. The first two are operated in an unsupervised fashion to segment meeting dialogues. For instance, Galley et al. (2003) extended the lexical-based TextTiling approach (i.e., LCSeg); Purver et al. (2006) adapted the sequence labeling approach, using topic models generated from other domains.

Galley et al. (2003) also applied supervised learning to classify segment boundaries, using both the outputs from LCSeg, which indicated information cohesiveness, and the other pragmatic features, which indicated speaker intentions. Results show that the classification approach which included non-verbal features (e.g., overlap rate, pause, and speaker change) outperforms the word-only approach. In addition, Banerjee and Rudnicky (2007) pointed out that the performance can be further improved when more distinctive participant interaction cues are aggregated into the segmentation model, such as the timing of note-taking.

The superior performance of the approaches that incorporate more multimodal features into the supervised segmentation model suggests the benefits of such incorporation. However, the effectiveness of the supervised learning approach has not be thoroughly examined: First, it is undetermined whether the proposed segmenter will perform as well on the task of finding hierarchical discourse structure of meeting dialogues, a task that is similar to the hierarchical topic detection task in texts or in read speech such as broadcast news (Trieschnigg and Kraaij, 2005). Few of the previously proposed models have considered the hierarchical nature of dialogue. One exception is Chino and Tsuboi (1996), who proposed an exchange structure (EX)-based discourse structure model for spontaneous two-party dialogues such as telephone conversation. In this model, a main discourse segment is defined as a parent node of a series of EXs that are initiated to realize a goal. However, this model is proposed for the context of

Japanese, and to the best of our knowledge no further computational research has been done using this model.

Current work in meeting dialogue segmentation has largely focused on ICSI meeting (LDC2004S02) segmentation (Galley et al., 2003; Purver et al., 2006), which focused on findings a single layer of topic segments that are indicated by words. (ICSI meeting segmentation is further characterized and analyzed in Chapter 6.) Because we expect meeting discourse segments to be indicated also by meeting activities (e.g., presentation, agenda discussion), further experiments are needed to determine whether the algorithms for finding topic segmentation may be able to find the activity-based segments.

Second, it is undetermined how the proposed segmenter will be affected by operating on imperfect transcripts generated by automatic speech recognizers (ASR). Since we cannot expect perfect human transcripts will always be available, the impact of using ASR output as opposed to human transcription on the performance of the segmenter has to be investigated.

Lastly, it is undetermined whether the multimodal and multiparty features can further help improve performance. In much other segmentation research (as discussed in Section 2.5), a variety of features have been identified as indicative of segment boundaries in different types of recorded speech. For example, Brown et al. (1980) showed that a discourse segment often starts with relatively high pitched sounds and ends with sounds of pitch within a more compressed range. Passonneau and Litman (1993) identified that topic shifts often occur after a pause of relatively long duration. Other prosodic cues (e.g., pitch contour, energy) have been studied for their correlation with story segments in read speech (Tur et al., 2001; Levow, 2004; Christensen et al., 2005) and with theory-based discourse segments in spontaneous speech (e.g., direction-giving monologue) (Hirschberg and Nakatani, 1996). In addition head and hand/forearm movements are used to detect group-action based segments (McCowan et al., 2005; Al-Hames et al., 2005).

In summary, in this thesis, we will further extend previous work to (1) address the problem of hierarchical discourse segmentation, (2) examine the effect of automatically generated features on segmentation performance, and (3) combine additional features that can be extracted from the pragmatic contexts of discourse units. In particular, the analysis conducted in this thesis provides a quantitative account of the effect of the verbal and non-verbal features on discourse segmentation. This is expected to lend support to the development of natural dialogue understanding applications.

2.6.2 Dialogue act classification

Answer	answering a question
Understanding	whether the speaker understood previous dialogue act
Signal-non-under	speaker did not understand
Signal-under	speaker did understand
Acknowledgement	demonstrated via backchannel or assessment
Repeat-rephrase	demonstrated via repetition or reformulation
Completion	demonstrated via collaborative completion

Table 2.4: *DAMSL classification scheme.*

Since this thesis aims to develop natural dialogue understanding applications for meetings, in the next subsections we also review possible ways to derive speaker intents (in particular, those related to argument outcomes) from all available communication modalities. As a first step, we look for a proper speaker intent classification scheme that is generalizable to handle multimodal inputs.

Dialogue act tagging schemes have been commonly used in research on the classification of speaker intents in human natural dialogues. DAMSL (Dialogue Markup in Several Layers) and many of its derivatives have been proposed (Walker et al., 1997; Core and Allen, 1997). Table 2.4 provides an example DAMSL classification scheme. Many corpora have been annotated. For instance, the SwitchBoard corpus (Godfrey et al., 1992), which consists of bi-party telephone conversation speech recordings, were annotated with the SWBD-DAMSL tagset (Jurafsky et al., 1997). The ICSI Meeting corpora have been annotated with the Meeting Recorder Dialog Act (MRDA) tagset(Shriberg et al., 2004) and the Multidimensional Abstract Layered Tagset for Utterances (MALTUS) tagset (Popescu-Belis, 2004). The AMI Meeting corpus also has its own annotation scheme, which labels each dialogue act with one of the labels in Table 2.5[12].

Recently, researchers have attempted to train classifiers to automatically label dialogue acts. Ang et al. (2005) achieved 81.18% accuracy on predicting the ICSI MRDA class of the dialogue acts out of 5 classes, using both the word-based features extracted from manually segmented human transcriptions and the posterior probabilities yielded with decision trees. Compared to the chance accuracy (obtained by predicting as the

[12]See Section 3.1.1 for a more detailed description of the ICSI and AMI annotations.

Acts about information exchange	Inform, Elicit-Inform
Acts about possible actions	Suggest, Offer, Elicit-offer-or-suggestion
Commenting on previous discussion	Assess, Comment-about-understanding, Elicit-assessment, Elicit-comment-about-understanding
Social acts	Be-positive, Be-negative
Special classes to complete annotation	Backchannel, Stall, Fragment
All the others	Other

Table 2.5: *AMI DA classification scheme.*

majority class) as 55.08%, the MaxEnt classifier used in this study performed very well. Despite that using automatically segmented ASR transcriptions decreased the accuracy to 74.96%, the performance was still significantly better than the chance accuracy (57.07%). The performance level suggested that automatic recovery of speaker intention is achievable.

The task of automatic DA classification has also been viewed as a sequence decoding task in past research. Dynamic Bayesian Network (DBN)-based HMM approaches were used to tag dialogue acts in the ICSI and the AMI corpora by Ji and Bilmes (2005) and Dielmann and Renals (2007b) respectively. In the AMI meeting corpus, the best performing model achieves 57.8% accuracy (with a lenient measure). But note that due to the fact that the AMI meeting corpus contains longer conversations and a larger number of DA classes, the performance is not directly comparable.

2.6.3 Argument intent recovery

While automatic dialogue act classification (for understanding general speaker intention) has achieved some success, automatic derivation of argument intent is less studied. To identify the argument outcomes in meetings, formal models have been extended to accommodate various communication modalities and argumentative intents in meeting discourse. For example, Chaudhri et al. (2006) handcrafted the CALO ontology (as shown in Figure 2.2) to construct an extensive knowledge base of concepts pertaining to the users' office environment. Three layers are included: Communication, Convey, and Transmit. Communicate events in the Multi-Modal Discourse (MMD) ontology are driven by the Issue-Based Information System (IBIS) model (Kunz and

Ritte, 1970). Niekrasz et al. (2005) have augmented the CALO ontology to create MMD, adding three communication modalities, i.e., light (gesture), sound (verbal and non-verbal spoken communication), and ink (hand writing), into the physical medium through which the communication is transmitted.

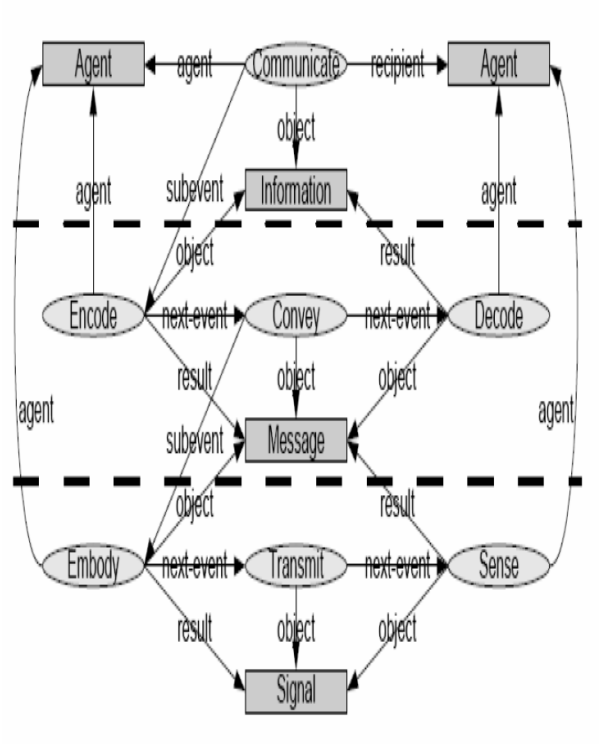

Figure 2.2: *CALO ontology. Events are depicted using ovals, entities using rectangles, relations using arrows.*

Further research has led to a number of new descriptive model proposals. For example, Marchand-Mailet (2003) proposed a classification scheme of argumentative acts (e.g., accept, request, reject) and a schema to organize and synchronize the acts. Pallotta et al. (2007b) extended the IBIS model to annotate the ICSI meeting dialogues. (An example argument diagram is shown in Figure 2.3 (adapted from Pallotta et al. (2007b).)

Rienks et al. (2005) further developed a new Twente Argument Scheme (TAS) to visualize the argumentative relations (i.e., positive, negative, uncertain) between the discourse units of different function roles (e.g., statement, open issue). An example diagram of an AMI meeting argument is shown in Figure 2.4.

Among the variety of models of meeting argument structures, only the TAS scheme has been attempted with automatic derivation. Rienks et al. (2005) reported the results

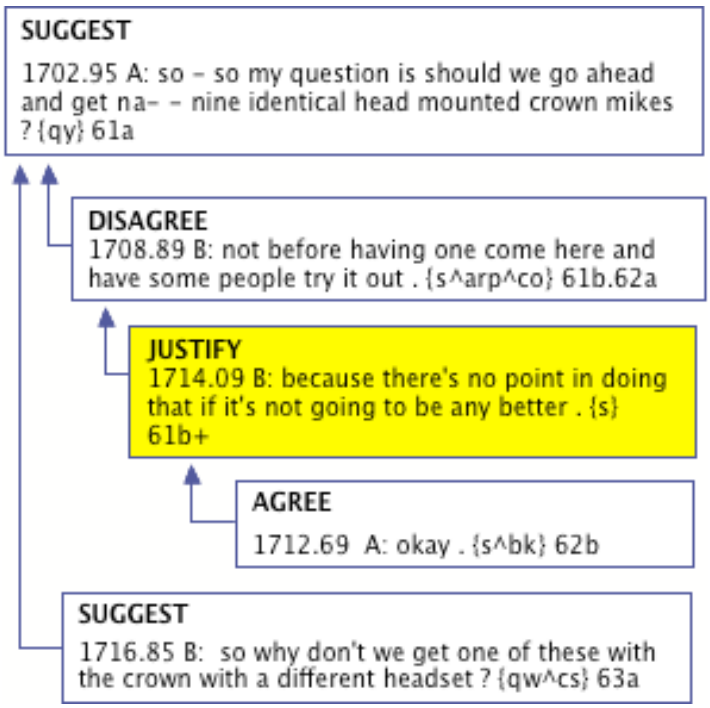

Figure 2.3: *Example argument diagram based on the IBIS model.*

of automatically classifying discourse units in the AMI scenario meeting corpus into 6 classes of argument acts and the linkage between units into 9 classes of argument relations. The best performing classifier achieved 65.88% and 58.14% accuracy with respect to the classification of argument acts and relations. The results also indicated that some of the argument act classes can be more accurately detected than others. For instance, positive argument acts (e.g., we agree, okay, true) can be detected with 77-79% accuracy. This has led to interest in combining the specific type of argument act detectors with the best performing discourse segmenters to "recognize important arguments along with their relevant contexts" – the capability we have envisioned in Section 1.2.

2.6.4 Meeting affect detection

Independent of speaker intent classification at the dialogue and argumentative act level, past research has also attempted to recognize high-level conversation phenomenon by including multiparty and multimodal features that can characterize the affective dimension (e.g., emotion, private state) of meeting speech. For example, paralinguistic features (e.g., spectral, and disfluencies) are used in Graciarena et al. (2006) to de-

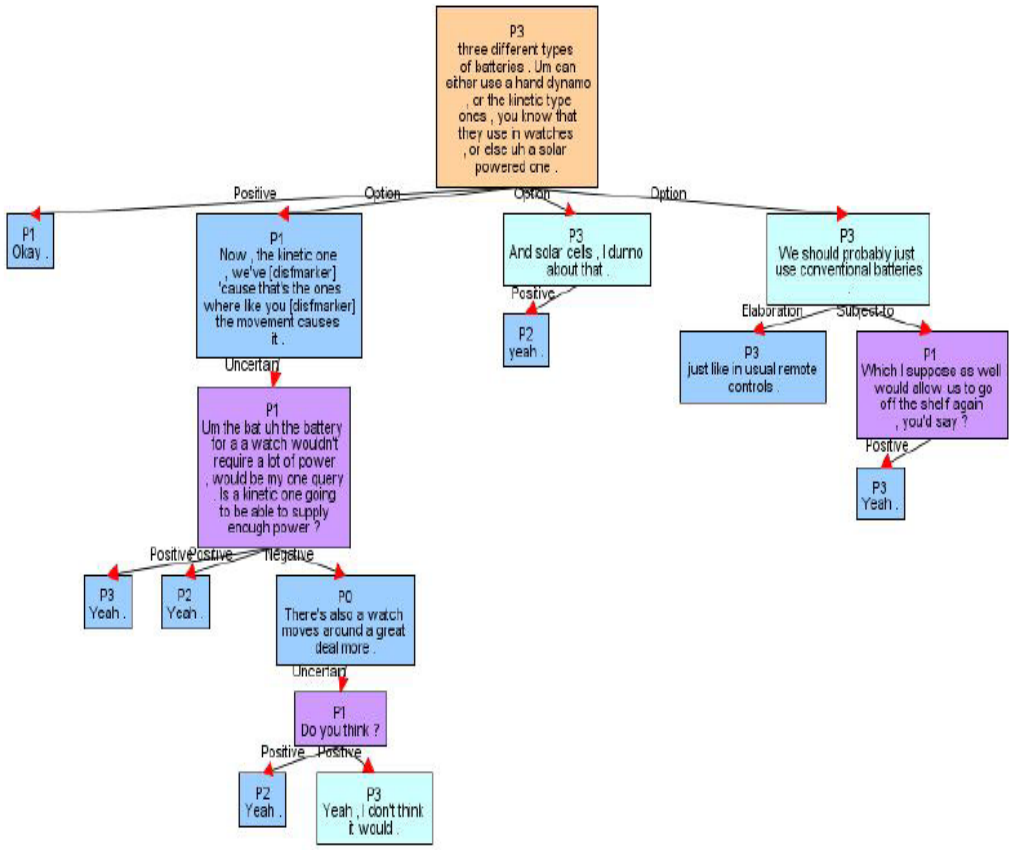

Figure 2.4: *Example TAS argument diagram.*

tect deceptive speech; pragmatic features (e.g., dialogue acts and adjacency pairs) are combined with prosodic features to detect agreement and disagreement in the argumentation process (Hillard et al., 2003; Galley et al., 2004) and hot spots (Wrede and Shriberg, 2003b,a); and video features (e.g., hand and head blob position) are combined to detect group-level activities (McCowan et al., 2005).

The use of affective-indicative features lends support to many other multiparty dialogue understanding applications, such as automatic detection of group-level activities (McCowan et al., 2005; Reiter and Rigoll, 2005; Zhang et al., 2005), participant roles and addressees (Banerjee and Rudnicky, 2006; Jovanovic et al., 2006), action items (Purver et al., 2006), and decision-related dialogue acts (Hsueh and Moore, 2007b; Fernndez et al., 2008). In Chapter 5, we will examine the effect of multimodal and multiparty interactive cues on the task of detecting decision-related meeting discussions.

2.6.5 Summarization

Extractive summaries can be used as the starting point of an information berry-picking process (Bates, 1990), in which each extracted sentence may trigger a follow-up action. Previous research has shown, when compared to full text, users can absorb information in summaries faster despite some loss of accuracy (Mani and Bloedorn, 1998). Previous research has also shown that the best automatic meeting summaries (see Erol et al. (2003)) are those that encapsulate answers to the most frequently asked questions.

Recent studies, in the domain of meeting speech summarization, have also shown a summary that is tailored to answer the most common query – what are the meeting decisions – to be critical to the preparation of future meetings (Pallotta et al., 2005; Rienks et al., 2005; Whittaker et al., 2005). Murray (2007) demonstrated that extractive summaries can help users achieve better performance in the "decision audit" task, in which users are asked to analyze the discussion of a particular decision made in a series of meetings. In these studies, the extractive summaries are displayed along with the transcripts and audio-video recordings in a meeting browser.

To save time and human labor in generating meeting minutes, recent research has developed automatic machinery to extract important utterances from the audio-video recording of spontaneous speech (Zechner, 2002; Christensen et al., 2003; Murray, 2007; Miekes et al., 2007). Traditionally, bottom-up, content-driven approaches are used to find generically important information. These approaches work well in information retrieval applications that support more exploratory types of search. The key challenge is to reasonably estimate the importance level of each sentence so that sentences can be extracted with respect to their rankings. This is often done by incorporating sentence-level (e.g., the position and title), lexical (e.g., the occurrence count of cue phrases, term frequency, co-occurrence, tf*idf score and its variants) (Edmundson, 1968; Kupiec et al., 1995; Teufel and Moens, 2002), and semantic features (e.g., the degree of connectedness in semantic graphs, co-reference) (Mani and Bloedorn, 1998; Barzilay, 2003). Past research in speech summarization has shown that, with the aid of prosodic features, the content-based approach can also work in a variety of speech genres widely ranging from broadcast news (Koumpis et al., 2001) to voicemail (Maskey and Hirschberg, 2005). Murray et al. (2005) have leveraged both the lexical and prosodic features to select the important utterances from meeting recordings.

Because the content-based approach does not satisfy the needs of applications that

are *query-driven*, more recently, approaches have been developed to tailor meeting extracts to a specific user focus. In addition, research in information extraction and question answering, as reviewed in Saggion et al. (2003), developed rule- and cue word-based approaches for extracting relevant information from text and speech in order to fill pre-determined templates. A related research area is the best automatic generation of meeting summaries (see Erol et al. (2003)) are one, that encapsulates answers to the most frequently asked questions.

However, whether rules and cues can be learned to fill a decision-based template is unknown. Therefore, in Chapter 5, we explore machine learning methods to do so by integrating multi-modal features (e.g., gesture and head movement) that are complementary to prosody and words. In addition, to the best of our knowledge, no studies have evaluated how effective the system- generated summaries are in helping users debrief the decisions made in a series of meetings. In Chapter 7, we will provide a quantitative account of how useful a decision query-based summary is to a real-life decision debriefing task.

2.7 Summary

In this chapter, we provided an overview of research in the area related to the recognition of discourse structure in meetings. In particular, we have discussed (1) the different notions of coherence central to discourse structuring, (2) the features which are characteristic of coherent units or coherent segment boundary sites, (3) the computational approaches effective for recovering discourse segmentation in a variety of speech genres including meetings, and (4) other possible ways to process meeting dialogues effectively.

The recovery of discourse segments requires continuous monitoring of lexical cohesiveness and other communication modalities, such as gesture/head movements, pitch, energy, rate of speech, and pause, which can indicate changes in pragmatic contexts. Although recent research that applied classification approaches has achieved some success in segmenting meetings, it has at least two shortcomings.

First, meeting dialogues are spontaneous conversations in a multiparty environment – naturally, we use more communicative modalities, including body language, gaze engagement, gesture, and prosody, to signal what we mean. There hence exists a large number of verbal and non-verbal features that may be indicative of speaker intentions. These features are expected to be complementary to one another. The correlation

among these features has yet been systematically studied.

Second, prior work has achieved some success in training classifiers to identify various conversational phenomenon in meeting dialogues; however, such approach requires plentiful labelled data. Therefore, in this thesis, we also look for unsupervised approaches. One of the most serious drawbacks of previous approaches using unsupervised methods is that they focus mainly on modeling lexical cohesiveness (as an indication of topical focus). Yet, previous research has also demonstrated the importance of incorporating features beyond words, such as multimodal and multiparty interaction cues, which are central to meeting speaker intention and structure recovery.

Thus begins our investigation into how to combine multiple knowledge sources into the unsupervised approaches. In order to do so, the following directions have been pursued: (1) a thorough empirical study of the systematic patterns in the verbal and non-verbal features (Chapter 4), (2) a quantitative account of the effect of the different combinations of features extracted from multiple communication modalities on meeting dialogue understanding (Chapter 5 and Chapter 6), and (3) novel ways to derive and combine features to enhance the performance of the existing approaches (Chapter 5 and Chapter 6).

Chapter 3

Corpus and Annotation

3.1 Meeting Corpus

Spontaneous face-to-face dialogues in meetings violate many assumptions made by techniques previously developed for broadcast news (e.g., TDT and TRECVID), telephone conversations (e.g., Switchboard) (Godfrey et al., 1992) , and human-computer dialogues (e.g., DARPA Communicator) (Eskenazi et al., 1999). In order to develop automated techniques to process multiparty dialogues, instrumented meeting rooms have been built at several institutes, and several corpora of meetings in natural contexts now exist, including those collected at the CMU Interactive Systems Labs (ISL) (Burger et al., 2002), the Linguistic Data Consortium (LDC) (Cieri et al., 2002), National Institute of Standards and Technology (NIST) (Garofolo et al., 2004), and the International Computer Science Institute (ICSI) at Berkeley (Janin et al., 2003). A subset of these corpora, an 80-minute and another 90-minute set, were used respectively in NIST's RichTranscription 2002 (RT-02) and 2004 (RT-04) Evaluation to examine the effectiveness of speech and video extraction technologies in the context of meeting dialogues (NIST, 2002, 2004).

Some more corpora were collected in the context of the CHIL "Computers in the Human Interaction Loop") project (Mostefa et al., 2008) and the IM2/M4 project (Marchand-Mailet, 2003). More recently, scenario-based meetings, in which participants are assigned to different roles and given specific tasks, have been recorded in the context of the CALO ("Cognitive Agent that Learns and Organizes") project (the Y2 Scenario Data) (CALO, 2006) and the AMI ("Augmented Multiparty Interaction") project (Carletta et al., 2006).

Table 3.1 presents the characteristics of these meeting corpora. These publicly

	CMU-ISL	LDC	NIST Pilot
Num. of meetings	60+ hrs 112 meetings	30 hrs 90 speakers	15 hrs 19 meetings, 72 speakers
Meeting type	natural meetings E.g., project planning, artificial meetings (scenario-driven)	natural meetings E.g., GroupMeet (with topics) GroupTalk (2-4 friends/family)	natural meetings E.g., party planning, artificial meetings, focus group
Evaluation	RichTranscription	RichTranscription	RichTranscription
	CHIL	CALO-Y2	ICSI
Num. of meetings	20 hrs 40 meetings	10 hrs 39 meetings	72 hrs 75 meetings, 53 speakers
Meeting type	natural meetings (in real-world) E.g., introduction, E.g., holiday planning	artificial meetings (scenario-driven) sets of 5 mtgs, 3-4 speakers) E.g., hiring	natural meetings (research lab meetings) E.g., speech recognition, E.g., meeting recording
Evaluation	RichTranscription CLEAR CLEF (Question-answering)	Role and Expertise detection	RichTranscription,
	IM2/M4	AMI	
Num. of meetings	39 short meetings	10 hrs	
Meeting type	artificial meetings (scripted) E.g., consensus, discussion disagreement	artificial meetings (scenario-driven) (sets of 4 mtgs, 4 speakers) E.g., product design (intro, conceptual/detail design, wrap up)	
Evaluation		RichTranscription CLEAR CLEF (Question answering)	

Table 3.1: *Meeting corpora.*

available recordings have been used to develop algorithms for a wider range of natural dialogue understanding applications, such as speech recognition, speaker turn segmentation, keyword recognition, speaker identification and tracking, recognition of emotional content, recognition of gestures and meeting activities, recognition of participant focus-of-attention, dialogue act segmentation, meeting phase segmentation, discourse segmentation, summarisation, and keyword search. These corpora have served as the foundation of formal evaluations, such as those conducted in the NIST-sponsored RichTranscription workshop, the CLEAR (Classification of Events, Activities and Relationships) evaluation, and the European Language Resource Association (ELRA)-sponsored Cross-Language Evaluation Forum (CLEF). These formal evaluations aim to gauge the progress made in speech-to-text transcription and metadata extraction techniques for conversational speech and techniques for tracking objects and hand pose in videos and classifying acoustic events in audio recordings.

3.1.1 Corpora used: the ICSI and AMI meeting corpus

In this thesis, I mainly use the AMI meeting corpus which consists of audio-video recordings of 173 meetings collected across three sites, the IDIAP research institute[1], University of Edinburgh and TNO[2]. This corpus also includes high quality, manually produced orthographic transcription for each individual speaker.

As the annotators at the Edinburgh site have also annotated hierarchical discourse segmentation on the ICSI meeting corpus, I also use the ICSI meetings for the empirical analysis of discourse segments and the development of automatic segmentation techniques in meetings. The ICSI meeting corpus (LDC2004S02) consists of the audio recording of seventy-five natural meetings in ICSI research groups. These meetings were recorded using close-talking head-mounted microphones and four desktop PZM microphones. The corpus includes manual orthographic transcriptions of all 75 meetings.

The AMI and the ICSI meetings are different in several respects. First, while all of the ICSI meetings are natural group meetings where participants met for their own real-life purposes, approximately two-thirds of the AMI meetings (140 out of 173) are driven by a scenario. In the scenario, four participants play the roles of project manager, marketing expert, industrial designer, and user interface designer in a design team, taking a design project from kick-off to completion. Second, while the

[1] http://idiap.ch/about.php
[2] http://www.tno.nl/home.cfm

ICSI meeting corpus only contains audio recordings, the AMI meeting corpus includes many other types of multimodal input, including close-up videos of participants, wide-view videos of the room, images from the projector, content of the whiteboard, and content of participants' hand-written notes. Third, the AMI scenario meetings follows a certain structure predetermined in the meeting agenda, while the ICSI meetings are less structured.

3.1.2 Transcription and ASR

In the AMI project, the entire AMI and ICSI meeting corpora were transcribed. The procedure is as follows: The speech signal has been automatically segmented into speech and non-speech first. The transcribers were asked to check and adjust the boundaries of the segments that contained speech such that the breakpoints reflected natural linguistic points in the utterance. Words and vocal noises (e.g., laugh, cough) were then tagged to reflect what the speakers have said.

In addition to the manual transcriptions, the AMI project team also generated ASR transcriptions for both the AMI and ICSI meeting corpora. The ASR transcriptions were produced by Hain et al. (2005), with an average word error rate (WER) of roughly 30%.

The system used a vocabulary of 50,000 words, together with a trigram language model trained on a combination of in-domain meeting data, related texts found by web search, conversational telephone speech (CTS) transcripts and broadcast news transcripts (about 10^9 words in total), resulting in a test-set perplexity of approximately 80. The acoustic models comprised a set of context-dependent hidden Markov models, using gaussian mixture model output distributions. These were initially trained on CTS acoustic training data, and were adapted to the ICSI meetings domain using maximum a posteriori (MAP) adaptation. Further adaptation to individual speakers was achieved using vocal tract length normalization and maximum likelihood linear regression.

A four-fold cross-validation technique was employed: four recognizers were trained, with each employing 75% of the meeting corpus as acoustic and language model training data, and then used to recognize the remaining 25% of the meetings. Table 3.2 summarizes the detailed statistics of the system word error rates.

	AMI Substitution	AMI Deletion	AMI Insertion	AMI TOTAL (WER)	ICSI TOTAL (WER)
Error Rate	21.3%	6.8%	4.1%	31.7%	25.3%

Table 3.2: *Word error rates of the ASR system in the AMI and ICSI corpus.*

3.2 Multi-Layer Annotation

To identify the features that can characterize decision decision discussions and their relevant contexts, in Chapter 4 I will perform an empirical analysis on a number of features that are expected to be characteristic and can be extracted from the neighboring context of decision discussions in the meeting corpus. In the following section, I will introduce the four kinds of annotations that are going to be used in the analysis: transcription, hierarchical discourse segmentation, dialogue act classification, and extractive and abstractive summaries.

In addition, to examine the actual effect of features on the task of automatic decision detection and discourse segmentation, in Chapter 5 and Chapter 6 I will perform a series of experiments to train models from the features that have been identified as potentially useful. For those features that perform well in the experiments, eventually we would like to generate them in an automatic fashion. Therefore, in the following section I do not just describe the manual procedure the AMI team used to produce the annotations needed for extracting these features, but also the automatic generation procedure of these annotations. The error rate of the automatically annotated features would then facilitate further discussion in the possibility to fully automate the process.

3.2.1 Hierarchical discourse segmentation

3.2.1.1 Discourse segmentation annotation

The AMI project team has produced discourse segmentation annotations for both the ICSI and AMI corpus. Although one third of the ICSI meeting corpus (25 out of 75) comes with annotations of discourse segmentation, I do not use these annotations in the empirical analysis and experiments. In Section 3.2.1.3 I will compare the inter-coder agreement obtained from the Edinburgh annotations of discourse segmentation on the ICSI meetings and with that obtained from the original ICSI annotations. Three

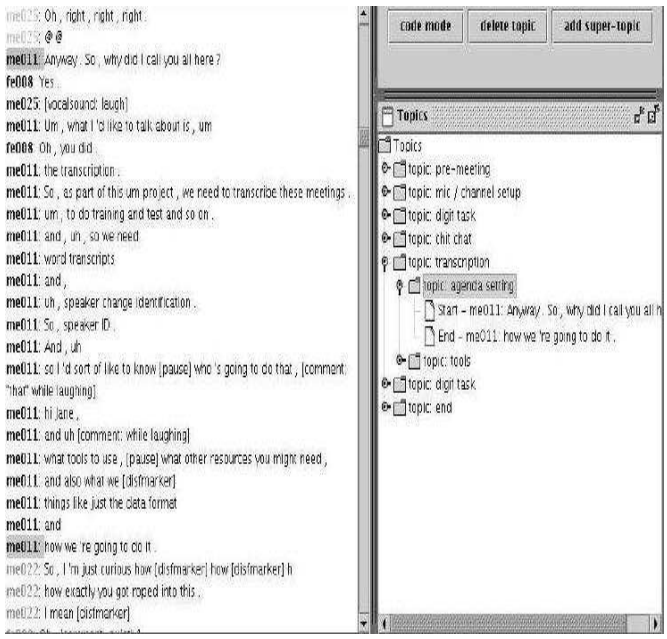

Figure 3.1: *The NXT tool is used to facilitate the multimodal annotation work. It contains built-in tools for media sync and data analysis and allows other customized plug-ins. In this example, the plug-in on the right hand side shows the topic structure of the meeting.*

human annotators used a tailored NXT annotation tool (as shown in Figure 3.1[3]) to perform discourse segmentation in which they could choose to decompose a topic into subtopics, with at most three levels in the resulting hierarchy. Following this procedure, a complete manual discourse segmentation has been annotated for the entire ICSI and AMI corpus.

It is expected that the preferred segmentation algorithm for predicting segment boundaries at different levels of granularity given different user tasks. Both the ICSI and the AMI project took an application-driven approach to determine preferred segmentation. That is, if the users were reviewing a meeting they might not have attended, what segmentation would help them quickly "drill down" to portions they might be particularly interested in reviewing. The AMI annotations were given the freedom to mark down as many hierarchies in the segmentation as possbile. However, even with the freedom, the annotators tended to stop at two levels of granularity. Therefore, in the following sections we explore segmentation at two levels of granularity.

We flattened the subtopic structure and considered only two leve ls of segmentation: top-level segments (TOPSEG) and all segments including subdiscourse segments

[3]This annotation tool is an open-source tool developed in Edinburgh for annotating multimodal corpora.

(ALLSEG). The top level of the structure signals either major topic shifts in discourse structure or serious disruption of the ongoing discussion. The second level of the structure signifies either a temporary digression or a discussion that is more focused on one aspect of the current major topic.

Basic statistics of the discourse segmentation annotations are reported in Table 3.3. Compared to the ICSI corpus, the segmentation structure of the AMI corpus is shallower, yielding smaller differences between the number of TOPSEG segments and that of ALLSEG segments.

- opening
- general discourse features for higher layers
- how to proceed
  - -+ segmenting off regions of features
  - -+ ad-hoc probabilities
  - -+ data collection
  - -+ experimental setup
- closing

Figure 3.2: *A list of major discussions (-) and sub-segments (-+) in the meeting Bed003 of the ICSI corpus.*

For example, a 60-minute meeting $\langle Bed003 \rangle$ in the ICSI corpus can be segmented into the hierarchical form as shown in Figure 3.2. In this meeting, the research team is discussing the planning of an automatic speech recognition project. Four major topics have been brought up: "opening", "general discourse features for higher layers", "how to proceed" and "closing". Depending on the complexity, each discussion of these topics can be further divided into a number of subdiscourse segments. For instance, the discussion of "how to proceed" can be subdivided to 4 subdiscourse segments: "segmenting off regions of features", "ad-hoc probabilities", "data collection" and "experimental setup".

3.2.1.2 Segment description annotation

Some differences exist between the annotations of segment descriptions in the two corpora. For the ICSI meetings, segment descriptions were essentially free format. Annotators were asked to provide a free text description for each discourse segment.

Average	TOPSEG	ALLSEG	Length
ICSI	6.96	17.2	40 mins
AMI	7.67	13.65	28 mins

Table 3.3: *Basic statistics of discourse segmentation annotations in the ICSI and the AMI corpus. ALLSEG segments refer to the combination of top-level and sub-level segments.*

They were encouraged to use keywords drawn from the transcription in the descriptions. These descriptions were saved in a database of labels that were accessible by all annotators; To impose some level of consistency, the annotators were also encouraged to select segment descriptions from the label database. For annotating the off-topic discussions, such as "opening" and "chitchat", some standard labels were also provided.

Because AMI meetings are scenario-based and include an agenda, it is expected that many of the discourse segments would correspond to those on the agenda. Annotators were given a set of segment descriptions to be used as labels. These labels are not just for the off-topic segments, but also for the other segments in which meeting participants are discussing scenario-related topics. Annotators were instructed to add a new label only if they could not find a match in the standard set.

The set of segment descriptions in the database can be roughly divided to three categories with respect to the the types of main topics discussed in the segments: **activity-based** (e.g., "*presentation*", "*discussion*"), **issue-based** (e.g., "*budget*", "*usability*"), and **off-topic functional segments (FUNC)** which serve the purpose of smoothing the procession of a discussion rather than that of contributing to the discussion (e.g., "*chitchat*", "*opening*", "*closing*", "*agenda/issues on recording equipments*").

We have observed that annotators tended to annotate activity-based segments only at the top level, whereas they often included sub-topics when finding **issue-based** segments. For example, a top-level **activity-based** segment, "*interface specialist presentation*", can be divided into four subdiscourse segments: "*agenda/equipment issues*", "*user requirements*", "*existing products*", and "*look and usability*".

In this two-layer structure, the functional segments (FUNC) account for approximately 42% of the top-level segments and 26% of all segments. The functional segments on average last around only one minute. Since meeting participants are expected to talk and interact differently during off-topic discussions, in the analysis reported in the following chapters, I will also analyze the characteristics of this type of segments. Table 3.4 presents the basic statistics of the following four types of discourse seg-

ments.[4]

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

	ALLSEG	TOPSEG	SUBSEG	FUNC
Average number of segments per meeting (± Stddev)	13.65 (± 8.16)	7.67 (± 2.42)	7.05 (± 6.68)	3.54 (± 2.27)
Average duration per meeting (in minutes) (± Stddev)	2.85 (± 3.23)	3.55 (± 3.84)	1.94 (± 1.86)	1.05 (± 0.89)

Table 3.4: *Basic statistics of the discourse segmentation annotations in the AMI corpora.*

3.2.1.3 Reliability

Previous research has examined the reliability of human discourse segmentation annotations under different definitions of discourse segmentation. For example, under the Rhetorical Structure Theory (RST) put forth in Mann and Thompson (1988), discourse segments are viewed as where the text that serves for one rhetorical relation switches to another. For example, from motivation to evidence to justification. Annotation studies showed that human annotators largely agreed with each other on where the rhetorical relation switches within a margin of a few utterances.

Passonneau and Litman (1993) demonstrated the level of reliability of human segmentation annotations remained within a reasonable range in spoken narratives. But

[4]Note that the number of all segments is not equal to the sum of top-level and subsegments. This is because there are a few annotator-specified subsegment boundaries that overlap with their top-level segment boundaries.

they also reported that it is impractical for naive subjects to annotate hierarchical segments longer than 200 words[5]

However, can human annotators agree on how to segment the often-lengthy meeting dialogues that involve interactions among multiple parties? Gruenstein et al. (2005) attempted to assess the reliability of the ICSI segmentation annotation procedure described in Galley et al. (2003), in which the ground truth was constructed by selecting the majority codings that were agreed by at least three coders. As argued in Gruenstein et al. (2005), the procedure has achieved a reasonable level of reliability (as measured in Kappa) on the complex task of segmenting meting dialogues. Even though it did not reach the level commonly accepted in computational linguistics[6], Di Eugenio and Glass (2004) found that such interpretation does not hold true for all tasks.

To establish reliability of the AMI hierarchical segmentation annotation procedures, Kappa statistics κ (Carletta, 1996) were also calculated as a measurement of the agreement between the annotations of each pair of coders.[7] The AMI team first collected four annotations on two chosen meetings, $ES2008a$ and $ES2008b$[8]. Table 3.5 illustrates the level of the pair-wise inter-coder agreement obtained on these codings in the two AMI meetings. Coding 1, 2, 3, 4 refers to each of the four annotators.

We also calculated P_k and W_d scores, the conventional metrics for evaluating the dissimilarity of two segmentations, as an indicator of the intercoder disagreement rate. P_k (Beeferman et al., 1999) is the probability that two utterances drawn randomly from a document (in our case, a meeting transcript) are incorrectly identified as belonging to the same discourse segment. Window Diff (W_d) (Pevzner and Hearst, 2002) calculates the error rate by moving a sliding window across the meeting transcript counting the number of times the hypothesized and reference segment boundaries differ. The lower the P_k and W_d scores are, the better agreement the pair of annotators have achieved.

Table 3.6 shows the average kappa statistics of the three pairs of coders on the top-level and sub-level segmentation respectively. Note that Gruenstein et al. (2005) examined how reliable the annotations obtained in the ICSI site for the 25 meetings used in Galley et al. (2003). In this thesis we examined the annotations obtained in the

[5]Seven annotators worked on segmenting 20 narrative monologues, taken from Chafe (1980), about the same movie. The average length of the narratives was 700 words, and the participants found it difficult to annotate hierarchical segmentation.

[6]Kappa values over 0.67 are taken to indicate reliable inter-coder agreement in computational linguistics.

[7]$Kappa(\kappa = (Observedagreement - chanceagreement)/(1 - chanceagreement)$.

[8]We selected the two meetings from those recorded in the meeting room at the University of Edinburgh (ES series).

	ES2008a		ES2008b	
CODER PAIR	TOPSEG	ALLSEG	TOPSEG	ALLSEG
1,2	0.77	0.58	0.54	0.44
1,3	0.65	0.58	0.34	0.27
1,4	0.93	0.72	0.63	0.44
2,3	1.00	1.00	0.62	0.55
2,4	0.81	0.73	0.72	0.62
3,4	0.81	0.73	0.57	0.58
Average	0.83	0.72	0.57	0.48

Table 3.5: *Pair-wise inter-coder agreement of annotations at the TOPSEG and the ALLSEG segments in two AMI meetings.*

context of the AMI project for the whole set of 75 ICSI meetings.

The statistics reported in the first two rows (Row 1-2) were obtained from Gruenstein et al. (2005), in which the authors reported κ (P_k—W_d) of 0.41 (0.28—0.34) for determining the top-level and 0.45 (0.27—0.35) for all segments (including the sub-discourse ones). The annotations used in this study were obtained on the ICSI corpus based on the original ICSI annotation guideline.

Row 3-4 shows the statistics reported in Hsueh and Moore (2006), in which we reported that the AMI annotators achieved $\kappa = 0.79$ agreement on the TOPSEG boundaries and $\kappa = 0.73$ agreement on the ALLSEG boundaries in the ICSI corpus. The annotations used in this study were obtained on the ICSI corpus using the annotation guideline developed within the AMI project.

Both sites allowed the annotators to label the discourse segments freely. The major difference between the ICSI and AMI annotation guideline lies in that the AMI annotators were also given a set of possible topic descriptions they can select from as the labels of the identified discourse segments.

Compared to the level of agreement achieved by using the ICSI annotation procedure, $\kappa = 0.41$ (TOPSEG) and $\kappa = 0.45$ (SUBSEG), the level of agreement achieved by the AMI annotation procedure suggests a better replicability.

Finally, the statistics in Row 5-6 indicates that the AMI annotation procedure achieved κ (P_k—W_d) of 0.70 (0.11—0.17) on the TOPSEG boundaries and κ (P_k—W_d) of 0.60 (0.23—0.28) on the ALLSEG boundaries in the AMI corpus. The Kappa values can be used to argue for a reasonable level of agreement on the segment boundaries

in the AMI corpus. The P_k and W_d scores indicate a low level of intercoder disagreement among the codings of discourse segmentation (especially at the top level) in the AMI meeting corpus, further confirming the reliability of the AMI segmentation annotation procedure used in this thesis. To our knowledge the reported degree of intercoder agreement (and disagreement) is the best in the field of meeting dialogue segmentation.

Corpus	Annotators	Kappa	P_k	W_d
ICSI(TOPSEG) (Gruenstein et al., 2005)	ICSI	0.41	0.28	0.34
ICSI(ALLSEG) (Gruenstein et al., 2005)	ICSI	0.45	0.27	0.35
ICSI(TOPSEG) (Hsueh and Moore, 2006)	AMI	0.79	n/a	n/a
ICSI(ALLSEG) (Hsueh and Moore, 2006)	AMI	0.73	n/a	n/a
AMI (TOPSEG)	AMI	0.70	0.11	0.17
AMI (ALLSEG)	AMI	0.60	0.23	0.28

Table 3.6: *Average inter-coder agreement of annotations at the top-Level (TOPSEG) and those at both the top-level and the sub-Level segments (ALLSEG) in the ICSI and the AMI meetings.*

3.2.2 Dialogue act class labeling

DA Class	Description	Example DA Types	Percentage
Information	Giving and eliciting information	"Suggestion"	31.9%
Action	Making or eliciting suggestions or offers	"Elicit-suggestion'	9.8%
Commenting on the discussion	Making or eliciting assessments and comments about understanding	"Assessment"	22.6%
Segmentation	Not contributing to the content but allowing segmentation of the discourse	"Backchannel", "Stall", "Fragment"	31.8%
Other	A remainder class for utterances which convey an intention, but do not fit into the four previous categories	"Be-positive"	3.9%

Table 3.7: *Distribution of the major DA classes in the AMI corpus.*

Dialogue acts (DAs) are the minimal unit that is used to represent speaker intentions in meeting dialogues.[9] As described in Section 2.6.2, various dialogue act classi-

[9]The term "dialogue act" is a specialised extension of speech acts (Searle, 1969), the acts of the

fication schemes have been proposed, such as MRDA (Shriberg et al., 2004), MALTUS (Popescu-Belis, 2004), and the AMI dialogue act classification scheme. In this study, I use the AMI classification scheme, which is a dialogue act typology tailored for group decision-making.

The AMI DA annotation scheme consists of 15 dialogue act types, which can be organized into five major groups. Table 3.7 exhibits the property, the example DA types and the distribution of the five dialogue act classes in the AMI corpus.

The AMI scenario meetings contain, on average, around 800 DAs in each 30-minute recording. Table 3.8 presents the DA-related statistics of the different types of discourse segments.

	ALLSEG	TOPSEG	SUBSEG	FUNC
Average duration per meeting (in DAs)(± Stddev)	71.2 (± 80.67)	88.84 (± 95.92)	50.41 (± 1.86)	22.19 (± 0.89)

Table 3.8: *DA-related statistics of the discourse segmentation annotations in the AMI corpora.*

3.2.3 Extractive and abstractive summarisation

3.2.3.1 Annotating decision-related dialogue acts

It is difficult to determine whether a dialogue act contains information relevant to a decision without knowing what decisions have been made in the meeting. Therefore, in this study decision-related DAs were annotated in a three-phase process as shown in Figure 3.3:

- **Producing abstractive summaries**: For each meeting in the corpus, one group of annotators were first asked to produce abstractive summaries about what the group is working on, and the decisions, problems, and future actions discussed in the meeting. Annotators were instructed to produce these summaries for an absent project manager.

- **Producing extractive summaries**: Another group of annotators were asked to extract a set of dialogue acts (around 10%) that convey what the group is work-

speaker saying something with the intention of communicating with an audience. Usually the speaker means more than what she actually says, relying on the mutually shared background knowledge and the power of rationality and inference on the part of the audience.

ing on, decisions, problems, and future actions. In general terms the extraction reflects and supports what is in the abstractive summaries.

- **Annotating decision links**: After the annotators finished the extraction, they were asked to go though the extracted dialogue acts one by one, and judge whether they supported (in other words whether there is informational similarity with) any of the sentences in the four types of abstractive summaries (i.e., general abstract, decisions, problem, future actions). For example, if a dialogue act supports any sentence in the decision section of the abstractive summary, a "**decision link**" from the DA to the decision sentence in the abstractive summary is added.

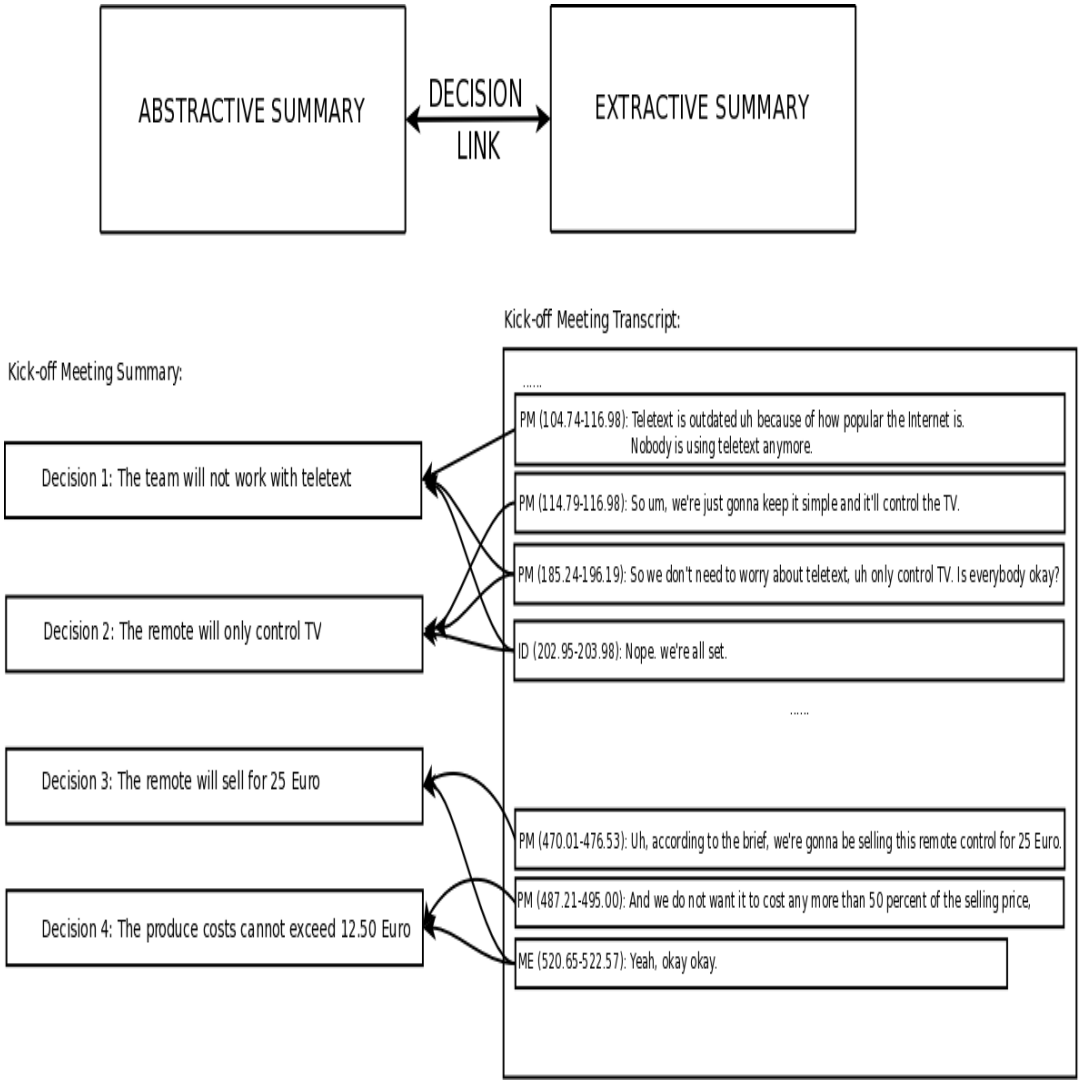

Figure 3.3: *Three-phase procedure for annotating decision-making DAs in the AMI corpus: abstractive summarization, extractive summarization, and decision linking.*

Following these annotation guidelines, a single extracted dialogue act can be linked with one or more sentences from the abstractive summaries. In some cases, it is possible that the annotator could not find any closely related sentence in the abstractive summary for the extracted dialogue act. In this scheme, it is not obligatory that there be a link. For those extracted DAs that do not have any closely related sentence in the abstract, the annotators were not obligated to specify a link.

Once the linking process is completed, we labeled the DAs that have one or more decision links as "decision-related DAs". In the 50 AMI meetings we used for analysis, the annotators found on average four decisions per meeting and specified around two decision links for each decision sentence in the abstractive summary. Overall, 554 out of 37,400 DAs were annotated as decision-related DAs, accounting for 1.4% of all DAs in the data set and 12.7% of the original extractive summaries (which consist of the extracted DAs).

3.2.3.2 Reliability

Among the three-phase procedure, the intercoder agreement of the extractive summarization phase was evaluated in Murray (2007). Murray reported the average Kappa value is 0.48 in the AMI meetings and argued that this level of agreement is common for summarization annotation.

To examine the level of inter-coder agreement of the decision-focused extractive summary, we calculated Kappa statistics κ between each pair of annotators' decision-focused extracts. Table 3.9 presents the reliability of the decision linking procedure.

Coder Pair	1,2	1,3	2,3	Average
Kappa	0.58	0.54	0.70	0.61

Table 3.9: *Pair-wise inter-coder agreement of decision-focused extract annotations.*

3.2.3.3 Annotating decision-related discourse segments

Decision-related discourse segments are operationalized as the discourse segments that contain one or more decision-related DAs. Overall, 198 out of 623 (31.78%) discourse segments in a set of 50 AMI meetings are decision-related segments. As the meetings we use are driven by a scenario, we expect to find that interlocutors are more likely to reach decisions when certain topics listed in a predetermined agenda are brought

up or when the discussions are related to the decisions made in previous meetings. For example, 80% of the segments labelled as "Costing" and 58% of those labelled "Budget" are decision-related discourse segments, whereas only 7% of the "Existing Product" segments and none of the "Trend-Watching" segments are decision-related discourse segments. (See Table 3.10 for a break-down of different types of decision-related segments.)

	ALLSEG	TOPSEG	SUBSEG	FUNC
Decision-related discourse segments per meeting	33%	31%	35%	4%
Decision-related dialogue acts per segment	3.7	4.5	2.76	3.83

Table 3.10: *Characteristics of discourse segments that contain decision-related DAs.*

Chapter 4

Towards Shallow Processing of Meeting Speech

4.1 Introduction

In Chapter 2, we provided a review of the features that have been used to predict spoken language phenomenon. In this chapter, we examine the use of these features, such as lexical cues and prosodic patterns, in detecting meeting decisions. In addition, as the AMI meetings are driven by a scenario, we also investigate the use of certain scenario-specific features, such as the speaker role, and other contextual features, such as the type of current dialogue act and its immediately preceding and following acts. One expected result out of this empirical analysis is the identification of the cue phrases and multimodal patterns that are predictive of decision-related discussions. In the end of this chapter, we will provide a summary of the identified decision-characteristic cues and patterns.

4.2 Analysis of Lexical Cues

Previous research has studied lexical differences (i.e., occurrence counts of N-grams) between various aspects of speech, such as discourse segmentation (Hsueh and Moore, 2006), speaker gender (Boulis and Ostendorf, 2005), and story-telling conversation (Gordon and Ganesan, 2005). As we expect that lexical differences also exist in decision-related conversations, we generated a decision-oriented language model from the decision-related Dialogue Acts in the corpus.

To compare the difference of decision-related conversations and general discus-

sions, we also generated a general language model from the rest of the conversations. Table 4.1 shows a list of the top n-grams in the general and the decision-oriented language models. Comparison of the two language models shows that some differences exist:

1. Decision making conversations are more likely to contain ***We*** than ***I*** or ***You***;
2. In decision-making conversations there are more explicit mentions of domain-specific content words, such as ***advanced chips*** and ***functional design***;
3. In decision-making conversations, there are fewer negative expressions, such as ***I don't think*** and ***I don't know***.

Unigram		Bigram		Trigram	
General	Decision	General	Decision	General	Decision
I	We	I think	I think	The remote control	The remote control
You	Uh	Yeah yeah	Remote control	I don't know	I think it's
Yeah	It	You know	We can	Yeah yeah yeah	We have to
Uh	So	Have to	You know	We have to	I think we
It	I	Remote control	We have	I think it's	Think we should
We	That	You can	So we	You have to	We need a
So	You	Kind of	The remote	Yeah I think	Maybe we can
That	Um	You have	You can	A lot of	The functional design
Um	Have	We have	Have to	I think that	The advanced chip
Have	Think	If you	We need	I think we	Then we can
Okay	Like	We can	We should	I don't think	On the side

Table 4.1: *Most frequent N-grams in the general language model and the decision-oriented language model.*

As shown in Table 4.1, the most frequent words (N-grams) are often "common words" which are not sufficient to describe a discussion. Thus, we need a measure that can quantify the discriminability of words to a particular type of discussion. In particular, we experimented with four different lexical discriminability measures – Log Likelihood ratio (LL), Chi-Squared statistics (X2), DICE coefficient (DICE), and Point-wise Mutual Information (PMI).

The LL and X2 measures capture the statistical association strength by summing over the amount of variation between the observed (O) and expected frequencies (E) in

the 2*x*2 contingency table to yield a single-valued parameter. For example, Table 4.2 shows a contingency table which exhibits how often the word "we" co-occurs with the decision-related DAs.

	in Decision DAs	not in Decision DAs	
"we"	212 (O_{11})	3,878 (O_{12})	4,090 (O_{1p})
not including "we"	5,805 (O_{21})	186,377 (O_{22})	192,182 (O_{2p})
	6,017 (O_{p1})	190,255 (O_{p2})	196,272 (O_{TOTAL})

Table 4.2: *Example 2x2 contingency table of word occurrences in decision DAs and non-decision ones.*

In the 2*x*2 contingency table, the values of the four cells correspond to the frequency of the given unigram (i.e., 'we'') occurring in the decision-related DAs (O_{11}) and that in all the other non-decision related DAs (O_{12}), and the frequency of all the other unigrams in the decision-related DAs ($O_{21} = O_{p1} - O_{11}$, where O_{p1} represents the total number of unigrams in the decision-related DAs) and that in the non-decision related DAs ($O_{22} = O_{p2} - O_{12}$, where O_{p2} represents the total number of unigrams in the non-decision related DAs). As shown in Equation 4.1, the expected frequency is computed as if the occurrences of the unigram "we" and that of the decision-related DAs were expected by chance.

$$E = O_{1p} \times (\frac{O_{p1}}{O_{TOTAL}}) = 4,090 \times (\frac{6,017}{196,272}) = 125 \tag{4.1}$$

Compared to the observed frequency of "we", the expected frequency is much higher:

$$O = O_{11} = 212 \tag{4.2}$$

It is posited in the LL and X2 measures that if a unigram occurs significantly more often in decision-related DAs than expected by chance, the unigram can be viewed as associated more strongly with decision-related discussions. In this case, the word "we" does occur significantly more often in decision-related DAs, and thus it is deemed a decision-discriminative word. The discriminability of this term can be further quantified using the LL and X2 measures as shown in Equation 4.3 and Equation 4.4.

$$X2(\text{"we"}) = 2 \times \sum(\frac{(O-E)}{E})^2 = 63.04 \tag{4.3}$$

$$LL(\text{"we"}) = 2 \times \sum(O \times \log(\frac{O}{E})) = 52.68 \tag{4.4}$$

Different from the statistical approach taken by LL and X2, the information theoretic PMI and DICE measures derive decision discriminability from the degree of mutual dependence of discrete events, i.e., the correlation coefficient between the occurrence of the target unigram and that of the decision-related DAs. As shown in Equation 4.5, the DICE coefficient is estimated by the observed frequency of the unigram in the decision-related DAs (O_{11}) and that in the whole training set (O_{1p}), and the total number of uni-grams in the decision-related DAs (O_{p1}). As shown in Equation 4.6, Point-wise Mutual Information (PMI) is defined as the log of the deviation between the observed frequency of a unigram in the decision-related DAs (O_{11}) and the expected frequency of this unigram if it were independent of the occurrence of decision-related DAs (E_{11}).

$$DICE(\text{“}we\text{”}) = \frac{(2O_{11})}{(O_{1p} + O_{p1})} = 0.042 \tag{4.5}$$

$$PMI(\text{“}we\text{”}) = \log(\frac{O_{11}}{E_{11}}) = 0.758 \tag{4.6}$$

Table 4.3 shows the list of discriminative words selected by the two categories of measures. (For space reason, we present only the bi-grams selected by the LL statistics and the DICE coefficient.) Comparing these N-grams with those in Table 4.1 shows that the discriminability helps to identify N-grams that are more characteristics of the decision-related conversations. A quantitative account of the effectiveness of these discriminability measures is provided in the experiment reported in Section 5.8.1.

As some of the statistically selected words are content words, the question of domain adaptability naturally emerges. While discrimnative common words can be carried over to detect decisions in another domain, content words tend to be domain-specific. A closer exmaniation of the top 10 words selected by the different statistical measures, there are 20-30% of domain-specific content words. If we further extend the list to the top 100 words, the content word ratio further increases to 50-60%. (See Appendix G for the list of the top 100 words selected by LogLikelihood.) The tendency suggests the necessity of retraining the models in another domain when supervised classifiers are used to develop decision detection models.

It is possible to lessen the annotation burden by applying transferring learning techniques. For example, for detecting speaker agreement and disagreement points, Hillard et al. (2003) employed an unsupervised clustering approach to label more unannotated words as agreement or disagreement markers. For detecting subjectivity in dccuments, Wiebe and Riloff (2005) trained a sentence-level subjectivity classifier from unanno-

LL	DICE
Advanced chip	Advanced chip
So no	Need a
Need a	So no
The wheel	We can
We can	As the
The LCD	The LCD
We're going	The wheel
Use a	We're going
On off	Remote control
We should	Use a
So we	On off
Remote control	So we
Like that	We should
We're gonna	We're gonna

Table 4.3: *Most discriminative lexical features (in bi-grams) selected by the measure of Log-likelihood (LL) statistics and DICE coefficient (DICE).*

tated data and aggregated the sentenc-level information to infer document-level subjectivity. Aue and Gamon (2005) used bootstrapping and meta-learning, e.g., applying the Expectation Maximization (EM) algorithm to train a generative classifier from out-of-domain data and then re-estimate the parameters of the classifier with unannotated in-domain data.

4.3 Analysis of Prosodic Patterns

In this study, we followed Shriberg and Stolcke (2001)'s direct modeling approach to extract the following prosodic features: **duration**, **pause**, **speech rate**, **pitch contour**, and **energy level**.

Duration: The minimal unit of this analysis is the dialogue act (**DA**). We reported the duration of each DA in term of both the number of words and the length in seconds.

Energy level: We used the normalized cross correlation function of the Snack Sound Toolkit to extract prosodic features from the sound files, computing a list of energy values delimited by frames of 10 ms. Then we applied a piecewise linearisation procedure to remove the outliers and average the linearised values of the units within

the time frame of a word.

Pitch contour: We also used the Snack Sound Toolkit to compute a list of pitch values and applied the linearisation procedure to obtain normalized **pitch values (PITCH)**. In addition, the pitch contour of a dialogue act is approximated by measuring the **pitch slope (SLOPE)** and **standard deviation (SD)** at multiple points within the dialogue act, e.g., the first and last 100 and 200 ms, the first, second, third and fourth quarter (Q1,Q2,Q3,Q4), the first and second half (H1,H2).

Speech rate: We used Festival's speech synthesis front-end to return phonemes and syllabification information. The rate of speech is then calculated as both the number of words spoken per second and the number of syllables per second.

These prosodic features are originally represented in the form of numrical values. However, the lexical features we analzed in the last subsection are represeted in the form of binary features, and to compare across differnt feature types, we need all features to be of the same type. Therefore, we first applied discretization to divide continuous features into discrete ranges. Discretization is the process of transforming continuous valued features to nominal and has been shown as an effective data preprocessing step for many machine learning algorithms (Quinlan, 1993; Almuallim and Dietterich, 1991).

The discriminability measures (i.e., LL, X2, PMI, DICE) were then applied to identify which discretized prosodic ranges of the prosodic features are most characteristics of decision-related conversations.

Table 4.4 summarizes the most discriminative prosodic features.

Functionally, prosodic features, i.e., energy, and fundamental frequency (F0), are indicative of segmentation and saliency. Prosodic structures have been shown as influential to the production and comprehension of syntactic analysis in dialogues (Cutler et al., 1997; Warren, 1999). On the ond hand, research in language comprehension has found many of the features important to resolving ambiguity either in read speech (Cooper et al., 1978; Lehiste, 1980) or in scripted dialogues (Schafer et al., 2004). On the other hand, research in language production has also confirmed that speakers do use prosodic differences to disambiguate, e.g., reducing the length of a description with repeated mentions (Clark and Schober, 1992), producing reduced form of repeated words in a discourse segment (Fowler, 1987).

Prosodic features have been identified as indicative of discourse structure in a variety of types of recorded speech. For example, Brown et al. (1980) and Menn and Boyce (1982) have shown that a discourse segment often starts with relatively high pitched

Feature Type	LL	DICE
Duration	$\|DA\|$ contains only one word	10 words $\leq \|DA\| <$ 14 words
(in words)	10 words $\leq \|DA\| <$ 14 words	
Duration	$\|DA\| \leq 1.5s$	$\|DA\| > 4.5s$
(in seconds)	4.5s $\leq \|DA\| < 5.4s$	
PITCH CONTOUR	PITCH remains stable	PITCH descends sharply in the end
Std. Deviation	$SD_{Q2} \simeq 0$	$SD_{Q1} \simeq 0$
	$SD_{H2} \simeq 0$	
Slope	$-0.5 \leq SLOPE_{H1} < 1$	$-8 \leq SLOPE_{Q4} < -1$
	$-0.5 \leq SLOPE_{Q1} < 1$	$-0.5 \leq SLOPE_{Q1} < 0$

Table 4.4: *Most discriminative prosodic features selected by Log-likelihood (LL) statistics and DICE coefficient (DICE). DA is the minimal unit used for prosody analysis. PITCH denotes the normalized pitch value. $SLOPE_{Qn}$ denotes the slope of pitch in the nth quarter of a DA. Likewise, $SLOPE_{Hn}$ denotes the slope of pitch in the nth half of a DA. SD_{Qn} denotes the standard deviation of pitch in the nth quarter of a DA.*

sounds and ends with sounds of pitch within a more compressed range. Passonneau and Litman (1993) identified that topic shifts often occur after a pause of relatively long duration. Other prosodic cues (e.g., pitch contour, energy) have been studied for their correlation with story segments in read speech (Tur et al., 2001; Levow, 2004; Christensen et al., 2005) and with theory-based discourse segments in spontaneous speech (e.g., direction-given monologue) (Hirschberg and Nakatani, 1996). Subjects can locate paragraph and sentence boundaries in conversational speech based only on prosodic cues (Kreiman, 1982).

However, as pointed out in Hastie et al. (2002), these identified correlations between prosodic features and discourse structure are often not a unique mapping. For example, dialogue acts that seek agreement are realized with both rising and falling boundary tones. Therefore stochastic models that assign a likelihood for each discourse structure functional role, such as that attempted by Wright (1998), cannot be trained with any single prosodic feature alone.

Likewise, previous research has found prosody useful for detecting the prominent words (or sentences) from long single-person speeches Arons (1994); Raux and Black (2003). The standard deviation and range of pitch are highly correlated with the emphasized (or stressed) portions of speech. Research of spontaneous speech has started receiving attention in the 90's. More pitch annotations in spontaneous speech were

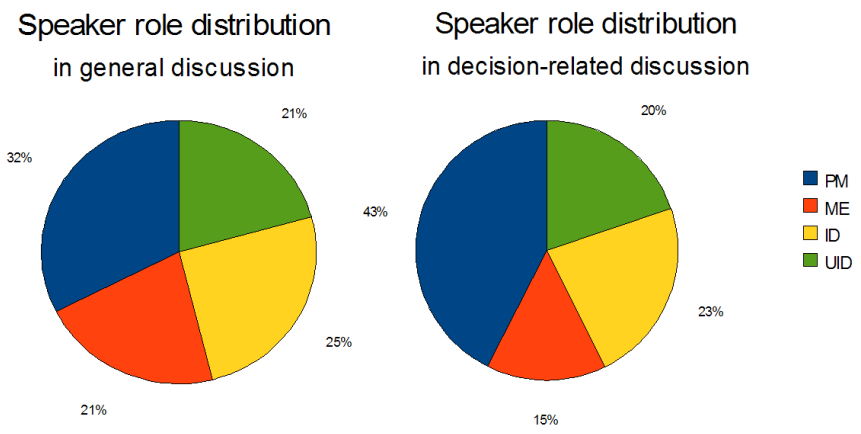

Figure 4.1: *Role distribution of speakers in general discussions (left) and decision speakers (right).*

made available (Wightman and Talkin, 1994). Kennedy and Ellis (2003) have shown that speakers in meetings often heighten pitch to highlight important information. This is consistent with our findings in Table 4.4 that heightened speech, which has its slope going up in the first half (or quarter) of sentence and going down in the second half (or quarter).

4.4 Analysis of Dialogue Context and Dialogue Acts

4.4.1 AMI-specific context

Contextual features specific to the AMI corpus, such as the speaker role (i.e., Product Manager (PM), Industrial Designer (ID), User Interface Designer (UI) and Marketing Expert (ME)), are expected to be characteristic of the decision-related DAs. Figure 4.1 shows the likelihood of participants in each role being a decision speaker. (In our definition, a decision speaker is the speaker of the decision-related dialogue act.)

In addition, the meeting type in a product design cycle (i.e., kick-off, conceptual design, functional design, evaluation) is also expected to be characteristic of the decision-related DAs. Figure 4.2 shows the likelihood of participants in each meeting type being a decision speaker.

Data analysis shows that the speaker role is a strong indicator of decision-related discussion, with 42.5% of the decision-related DAs generated by participants who played the role of PM. Wilcoxon Rank Sum Z-tests confirm that PMs significantly assumes a dominant role when meetings proceed to the decision-related discussion

($z = 4.59; p < 0.001$), and further find that MEs are significantly less likely to be a decision speaker ($z = -3.57; p < 0.001$).

In contrast, the meeting type is less indicative of decision-related discussion. The only difference that has been found as statistically significant is that meeting participants made relatively fewer decisions in the evaluation/wrap-up meetings ($z = -2.13; p < 0.05$).[1]

4.4.2 Decision-indicative dialogue act type

Our analysis also demonstrates that there is a statistically significant difference in the type of decision-related dialogue acts. Figure 4.3 demonstrates that meeting participants are more likely to use certain types of dialogue acts to express key decision-related information. As shown by the Z-test results in Table 4.5, dialogue acts in which speakers ***offer***, ***inform***, ***suggest***, or ***elicit assessment*** are more likely to be decision-related DAs. In contrast, DAs in which speakers ***elicit offer***, ***make comments***, ***elicit comments***, or simply ***signal understanding to keep the conversation flow*** are less likely to be decision-related.

Because meetings involve multiparty interactions, in this analysis we also examined the group behaviour-based properties of the decision-related DAs, such as whether the DA is a reflexive act and how many group members are this DA addressing to. A reflexive act is a dialogue act in which the group stepped back to discuss not the project

[1] This can be explained by the AMI meeting scenario: In the evaluation/wrap-up meeting, the participants were not asked to design any more new features, but to evaluate the prototype they have come up with in the previous three meetings.

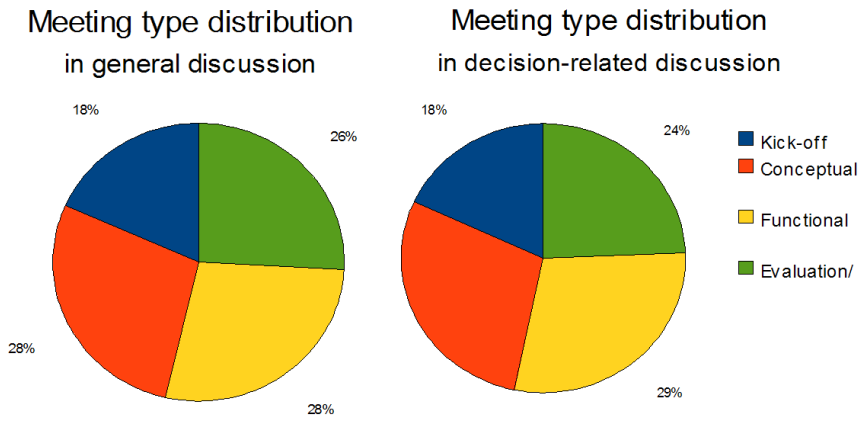

Figure 4.2: *Meeting type distribution of speakers in general discussions (left) and decision speakers (right).*

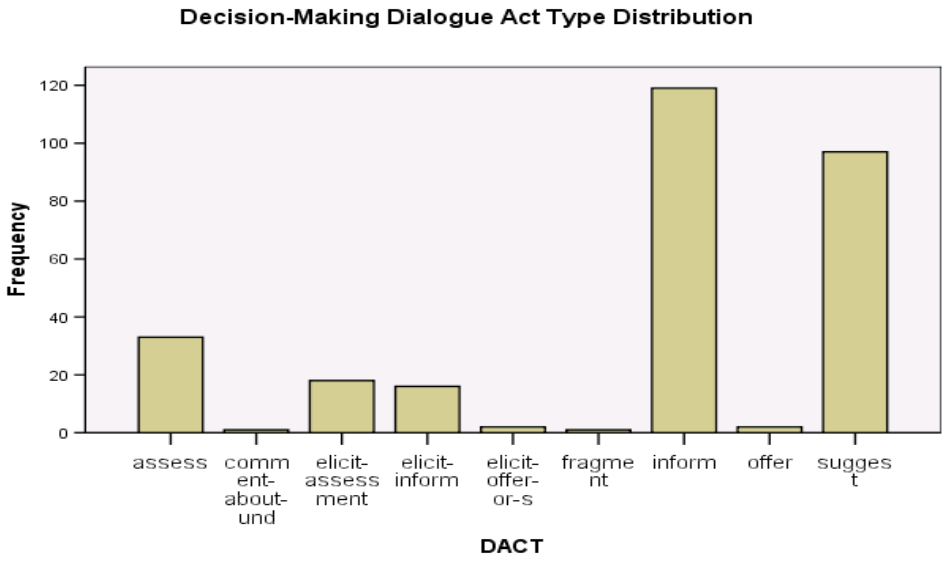

Figure 4.3: *Frequency of dialogue act types in key decision-related discussions.*

itself, but how they as a group were approaching the project. As shown in Table 4.6, when talking about decisions, the meeting participants tended to reflect more often on how the group was carrying out the task.

The analysis in Table 4.7 shows the decision-related DAs are identifiable by their immediately preceding or following DA. For example, a decision-related dialogue act is likely to be preceded by ***stalls***[2] or followed by ***fragments***[3] In addition, a decision-related DA is less likely to be preceded or followed by a ***suggest*** or other elicit type of DA (i.e., ***elicit-inform***, ***elicit-suggestion***, ***elicit-assessment***).

4.5 Analysis of Subjective Cues

Next we consider whether the decision-related information is expressed in a more subjective or objective manner. Previous research identified lexical items and phrases that are distinctive of subjective language in text (Turney, 2002; Cohen, 2003; Riloff and Wiebe, 2003; Yu and Hatzivassiloglou, 2003; Wilson et al., 2005). Do the speakers tend to use subjective language in decision-making conversations? Does there exist a difference between the subjective expressions in the spoken and written language?

[2]STALL is where people start talking before they are ready, or keep speaking when they haven't figured out what to say.

[3]FRAGMENT is the segment which is not really speech or is unclear enough to be transcribed, or where the speaker did not get far enough to express the intention.

DA TYPE(Significance Code)		Decision	General	Z-test
Acts about information exchange	Inform(***)	53.0%	28.0%	$z = 33.4$
	Elicit-Inform	3.9%	3.6%	$z = 0.9; NS$
Acts about possible actions	Suggest(***)	21.1%	8.2%	$z = 27.2$
	Offer(**)	1.8%	1.3%	$z = 2.9$
	Elicit-offer-or-suggestion(***)	1.1%	5.1%	$z = -31.4$
Commenting on previous discussion	Assess(***)	13.6%	17.0%	$z = -5.7$
	Comment(***)	0.3%	23.7%	$z = -93.6$
	Elicit-assessment(***)	4.5%	2.0%	$z = 10.2$
	Elicit-comment(**)	0.02%	0.26%	$z = -2.99$
Social acts	Be-negative	0.07%	0.11%	$z = -0.68; NS$
	Be-positive(**)	0.19%	2.01%	$z = -8.35$
Special classes to complete annotation	Backchannel(***)	0.02%	9.4%	$z = -20.7$
	Stall(***)	0.0%	8.3%	$z = -19.4$
	Fragment(***)	0.3%	14.1%	$z = -25.5$

Table 4.5: *DA type distribution of the AMI decision-related DAs and general DAs. Significance codes: (***):* $p < 0.001$*; (**):* $0.001 <= p < 0.01$*; (NS):* $p >= 0.05$.

Previous work has shown that the use of subjective language in written text does exhibit distinctive cues (Turney, 2002; Cohen, 2003; Riloff and Wiebe, 2003; Yu and Hatzivassiloglou, 2003; Wilson et al., 2005). In this work, we employed a list of 8,221 subjective terms (hereafter referred to as "MPQA Lexicon") that were previously collected from written texts (Wilson et al., 2006). These terms are categorized with regard to their subjective and sentiment properties. For example, each term is tagged with its degree of subjectivity (i.e., weak, strong) and the polarity of sentiment (i.e., positive, negative). Terms that do not come with an obvious positive or negative sentiment are tagged as "neutral". Terms that represent some level of polarity (either positive or negative) are tagged as "polar".

In addition, terms are also categorized with respect to their argument polarity, which indicates whether a word is positive or negative when used to express a belief or argument.[4]

To understand the distribution of subjective term types in meeting speech, we counted the number of terms in each subjective category that have occurred in our

[4]Note that the majority of subjective terms have a neutral arguing polarity, e.g., "wrath".

DA property(Significance Code)	Decision-related DA	General DA	Z-score
Reflexivity(***)	1.4%	9.2%	$z = -17.29$
Addresses to one(***)	22.5%	42.8%	$z = -25.42$
Addresses to two(***)	2.4%	1.0%	$z = 8.11$
Addresses to all(***)	75.1%	56.2%	$z = 23.57$

Table 4.6: *Difference in the level of reflexivity and the number of addresses between decision-related DAs and general DAs. Significance codes: (***): $p < 0.001$.*

DA TYPE		Preceding	Current	Following
Acts about information exchange	Inform	35.0%	41.2%	48.7%
	Elicit-Inform	1.7%	5.5%	3.3%
Acts about possible actions	Suggest	11.7%	33.6%	13.2%
	Offer	2.2%	0.7%	1.3%
	Elicit-offer-or-suggestion	1.1%	0.7%	0.7%
Commenting on previous discussion	Assess	14.4%	11.4%	12.5%
	Comment-about-understanding	2.2%	3.0%	1.3%
	Elicit-assessment	0.6%	6.2%	2.6%
Special classes to complete annotation	Backchannel	0.0%	0.0%	9.6%
	Stall	0.0%	0.0%	8.0%
	Fragment	3.0%	0.0%	14.1%

Table 4.7: *DA class distribution of the AMI DAs that immediately precede and follow the decision-related DAs.*

corpus. Table 4.8 presents the result. In order to assess whether the distribution in meeting speech is different that in text, in this table we also present the number of MPQA Lexicon terms in each subjective category. The ratio of the positive terms to the polar terms in our corpus, 73%, shows that meeting participants speak very positively. Comparing this ratio with that the MPQA Lexicon, which is only 33% indicates the difference in the chance of using positive terms in text and in meeting speech.

To examine whether the meeting participants have a strong tendency to use subjective language in decision-related discussions, we performed Z-tests to verify the hypothesis that there exists a difference between the proportion of subjective terms in the decision-related discussions and that in the general discussions. The Z-score signifies the statistical difference between the population that is sampled from the general

Type	Intensity	# of subj.	Polar	Positive	Negative	Neutral	All
Sentiment Cues	Strong	MPQA Lexicon	4,555	1,482	3,073	174	4,743
		AMI corpus	48	35	13	23	69
	Weak	MPQA Lexicon	1,934	842	1,092	257	2,188
		AMI corpus	60	36	24	33	93
	All	MPQA Lexicon	6,489	2,324	4,165	441	6,440
		AMI corpus	108	71	37	56	162
Arguing Cues	N/A	MPQA Lexicon	312	175	137	657	1,281
		AMI corpus	21	20	1	189	231

Table 4.8: *The number of subjective terms found in text and in meeting speech.*

discussions and the population that is sampled from the decision-related discussions. The results, shown in Table 4.9, indicate that such differences do exist and are statistically significant in all subjective categories except in the category of negative arguing terms ($z = -0.32; NS$).

A closer examination at the magnitude of the statistical significance suggests the following: (1) The subjective terms are used significantly more often in decision-related discussions ($z = 10.55; p < 0.001$). (2) Among the subjective terms of different polarity, the neutral terms is the most distinctive characteristics of the subjective language in decision-related discussions ($z = 19.71; p < 0.001$); (3) The arguing terms are also used significantly more often in decision-related discussions ($z = 8.82; p < 0.001$); (4) Among the arguing terms of different levels of polarity, the positive arguing terms have the most significantly different distribution ($z = 8.97; p < 0.001$).

In short, our corpus study show that the use of subjective language in meeting speech does exhibit demonstrable differences from that in text, and speakers do use subjective language significantly more often when it comes to decision time.

This analysis indicates that people do argue positively in the decision-related discussions, although they often choose to express it in a relatively neutral way. For example, people use the word "*think*" quite often when trying to reach a particular decision, and "*think*" is a word that is neutral in its sentiment polarity but positive in its arguing polarity.

CUE TYPE	SUBJ. INTENSITY	POSITIVE	NEGATIVE	NEUTRAL	ALL
	Strong	4.38(***)	4.42(***)	2.13(*)	3.15(***)
Sentiment Cues	Weak	8.04(***)	9.62(***)	12.27(***)	11.75(***)
	All	10.85(***)	10.65(***)	19.71(***)	10.55(***)
Arguing Cues	8.84(***)	8.97(***)	-0.32(NS)	5.29(***)	8.82(***)

Table 4.9: *Decision discriminability of the subjective terms in meeting speech. The value of decision discriminability in each cell represents the Z-score of the subjective terms in that particular subjective category. Significance codes: (***): $p < 0.001$; (*): $0.01 <= p < 0.05$; (NS): $p >= 0.05$.*

4.6 Summary

The analyses reported in this chapter show that **people do speak and interact differently** when discussing information critical to making decisions.

First, the analysis has identified the decision-characteristic lexical cues from the transcripts of the decision-critical discussions. For example, a decision is likely to be reached when the interlocutors use certain conventional expressions (e.g., *yeah*, *okay*, *mm*, *so*) and refer to certain terms or concepts. These include topical words (e.g., *usability*, *costing*, *discussion*, *evaluation*) and content words that are related to interface and industrial design (e.g., *buttons*, *advanced chip*, *slogan*, *teletext*) in the remote control design meetings of the AMI corpus. When talking about decision-critical information, the interlocutors address to the whole group (i.e. as "*we*") rather than just one or two group members. They also use more subjective language, especially positively arguing terms with a neutral sentiment (e.g., *think*).

Moreover, the analysis has identified other decision characteristic cues that are hidden in the audio-visual recordings. For example, the interlocutors often stress their points with a higher than usual pitch and with a long pause, so as to capture the other participants' attention.

Furthermore, the analysis has identified cues from the pragmatic context of meeting dialogues, for example, the speaker role. Decision-related discussions are dominated by PM.

In addition, the type of current dialogue act and its immediately preceding and following DA are cues to decision-making. We observed that decision-critical information is often expressed after participants have provided an evaluation of ideas (***assess***),

or suggested an action related to the group or another individual (***suggest***). Interlocutors often reveal decision-critical information when they are providing information (***inform***) or expressing an action-related intention (***elicit-suggest***).

The results reported in this chapter show systematic patterns in decision-related discussions that can be used to distinguish them from general discussion. The next question to be addressed is whether the systematic differences would be useful for decision recognition. To provide a qualitative account of the various decision-characteristic cues, in the next chapter we describe the experiments that were used to verify what set of features are really discriminative of decision-related DAs.

In this chapter, we also established a decision discriminability analysis framework, using the statistical measures (e.g., LogLikelihood ratio) and information theoretic measures (e.g., DICE coefficient). With the aid of this framework, the decision discriminability can be measured to rank the various types of cues. These cues are then later extracted for recognizing the decision-related discussions.

Chapter 5

Meeting Decision Detection

5.1 Introduction

To assist users in revisiting decisions within meeting archives, our goal is to automatically identify decision-related dialogue acts (decision DAs) and discourse segments (decision DSs). As the development of such an automatic decision detection component is critical to the re-use of meeting archives, it is expected to lend support to the development of other downstream applications, such as computer-assisted meeting tracking and understanding (e.g., assisting in the fulfilment of the decisions made in meetings) and group decision support systems (e.g., constructing group memory) (Post et al., 2004; Romano and Nunamaker, 2001).

Previous research has developed descriptive models of meeting discussions, focusing on modeling the dynamics of meetings (Niekrasz et al., 2005) or on modeling the content discussed in meetings (Marchand-Mailet, 2003; Rienks et al., 2005). While automatically extracting these argument models remains a challenging task, researchers have begun to make progress towards this goal (Galley et al., 2004; Gatica-Perez et al., 2005; Hillard et al., 2003; Hsueh and Moore, 2007b; Purver et al., 2006; Wrede and Shriberg, 2003b).

The goal of this chapter is to explore "automatic decision detection" in meeting speech, i.e. finding the sections of recordings that contain decision-related conversations, and to provide an interface that displays the decisions and the related conversations. In particular, this system focuses on locating decision-related information at two levels of granularity: discourse segments and dialogue acts. First, the system detects decision DAs that are both extract-worthy and reflective of the content of the decision discussions. Then, the system detects decision DSs, which we define as where meeting

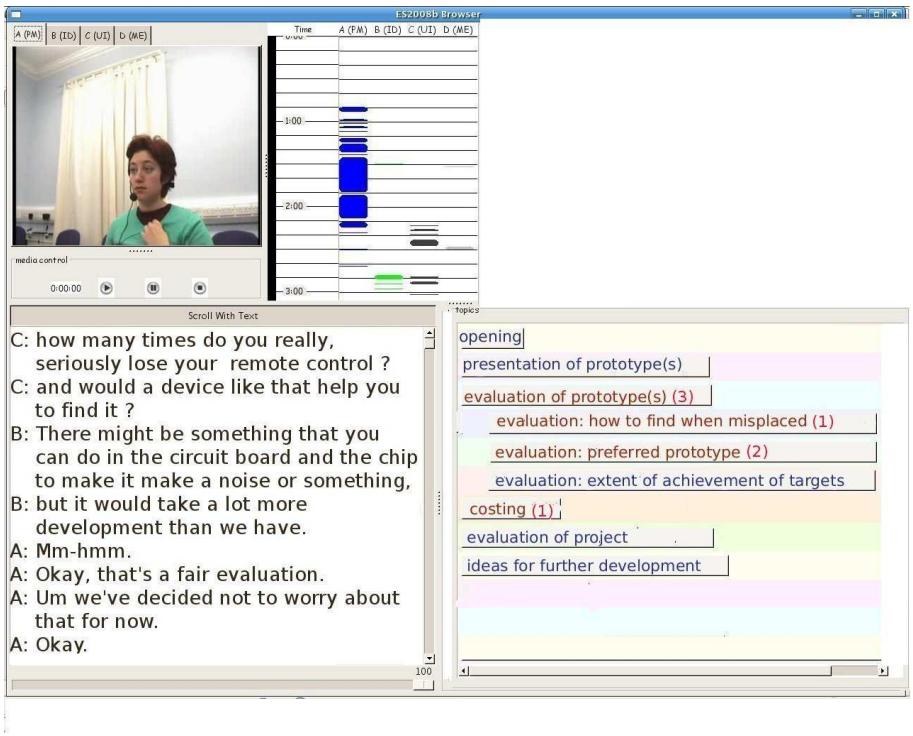

Figure 5.1: *Example application that demonstrates the use of decision DS information. The bottom right component shows a list of discourse segments in an example meeting. The discourse segments shaded in red are those that contain at least one decision. The number shown in parentheses following each segment label indicates the number of decisions reached within the segment.*

participants have reached at least one decision.

As shown in Figure 5.1, the detected decision DSs allow users to get an overview of the decisions made in previous meetings. (For a more detailed view of the displayed discourse segment labels, please see Figure 1.5.) After users have identified the decisions they are interested in, as shown in Figure 5.3, the detected decision DAs provide more details. For example, if a user spots an interesting discussion on "how to find (the remote) when misplaced", they can click on the segment button in Figure 5.1 that will take the user to the view of Figure 5.3. The user can then read through the excerpt of the discussion related to this decision and quickly find out the group has decided "not to worry about designing a function to find the remote when misplaced". (For a more detailed view of the excerpt, please see Figure 5.2.)

For the meeting extracts of a complete series of meetings, please refer to Appendix F.

PM: *But um the feature that we considered for it not getting lost*.
ME: *and we think that each of these are so distinctive, that it it's not just like another piece of technology around your house*.
ME: *So we're not thinking that it's gonna be as critical to have the loss*.
PM: *Um we so we do we've decided not to worry about that for now*.

Figure 5.2: *Example excerpt composed of decision DAs. The number in the parenthesis indicates the position of the selected DA in a discourse segment. In this example, annotators have selected four dialogue acts to represent the design decision of "how to find (the remote) when misplaced".*

In Chapter 3, we analyzed a number of features to determine whether they are discriminative of decision DAs. The analyses showed that people do "speak" differently when making decision, in terms of both word choices and expression styles.

For example, when reaching a decision, interlocutors are more likely to use certain conventional expressions (e.g., *yeah*, *okay*, *mm*, *so*), topical words (e.g., *usability*, *costing*, *discussion*, *evaluation*), and content words that are related to interface and industrial design (e.g., *buttons*, *advanced chip*, *slogan*, *teletext*) in the remote control design meetings of the AMI corpus.

As for expression style, when it comes to a decision point, interlocutors emphasize their points with higher than usual pitch and with longer than usual pauses before making the point, possibly to capture the others' attention. Speakers also address the whole group (i.e. as "*we*") rather than just one or two group members, and decision-related discussions are dominated by PM. In addition, interlocutors express a decision significantly more often when they are giving information (***inform***) or expressing an action-related intention (***suggest***). They usually make a decision after some participant has expressed an evaluation (***assess***), an intention related to the actions of the group or another individual (***suggest***), or an intention related to their own actions (***offer***). Finally, interlocutors use more subjective language, especially the positively arguing terms (e.g., *think*),

To capture the above decision characteristics and develop an automatic meeting decision detection (MDD) system, the following four types of models were created: decision-specific language model (LX), prosody model (PROS), dialogue act model (DA), and topic model (TOPIC). The first experiment of this chapter combines information from all four of the models to perform decision detection.

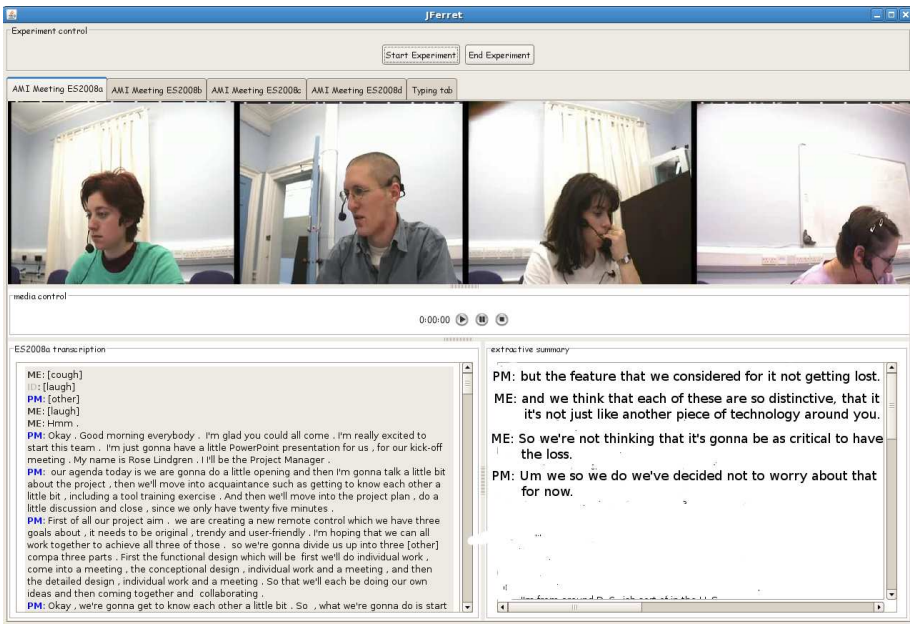

Figure 5.3: *Example application that demonstrates the use of decision DA information. The bottom right component shows a set of decision DA extracts that are representative of the design decision of "how to find (the remote) when misplaced".*

Note that Experiment 1 uses only the dialogue acts that are included in the extractive summaries. This is because our annotators examined only the DAs in the extractive summaries for decision links (c.f. 3.2.3.1 for annotation procedure). As we wish to compare the automatically detected results fairly with the manually selected ones, the automatic algorithm must operate on the same input. Also, the task of automatically producing extractive summaries for the AMI meetings has been attempted with the state-of-the-art algorithm, yielding 21%-22% accuracy (F1 score) in Murray (2007). However, as we eventually would like to identify decision DAs directly from the transcripts, in Experiment 2, the decision detection component is also operated on the entire transcript.

In addition, many of the cues that have been identified as characteristic of decision DAs in Chapter 4 are manual features, i.e., they rely on the existence of manual annotations, such as human transcription and DA classification. Because these manual annotations are not always available, we also need to evaluate whether the performance of MDD will be seriously degraded by replacing the manual features with their automatically generated versions. In Experiment 3, we report the results of training the decision detection component with features extracted from transcriptions produced by

an automatic speech recognition (ASR) system and automatically annotated dialogue acts.

We also explore the use of subjective term features in automatic decision detection. As the use of subjective language was shown to be characteristic of decision-related conversations, in Experiment 4, we explore the impact of using the subjective term features on the accuracy of the decision prediction models.

Finally, since not all of the cues that are extracted from the audio-video recordings and the meeting transcripts contribute equally to the accuracy of predictions, in Experiment 5 we experiment with various methods to select the most decision characteristic cues in the early fusion stage and examine the impact of incorporating these methods on the accuracy of the automatic decision detection component. The feature selection methods we explore are those that have been shown to be useful in text classification. In particular, we use two statistical and two information theoretic measures to determine the association strength between the occurrence of a particular decision-characteristic cue and that of the decision DAs.

Having selected the predictive cues, in Experiment 6 we explore the use of ensemble modeling techniques, which leverage the prediction results of various modalities and feature selection methods in yielding a final prediction. We evaluate the impact of these techniques in a late fusion stage on improving the accuracy of predictions.

In sum, I will report the results of these experiments and answer the following questions:

1. Can automatic machinery be developed to detect decision-related conversations by integrating the potentially characteristic but widely ranging features, e.g., word choices, multimodal cues? Can the decision detection component can be improved by monitoring the use of subjective language? And if so, what are the most discriminative knowledge sources for the decision detection task?

2. Can the decision detection component be operated fully automatically? Will its performance seriously degrade when used to detect decisions directly from complete recordings? Is any manual annotation necessary to the development of a well-performing decision detection component?

3. How can we integrate multiple knowledge sources most effectively? How does the performance of the automatic decision detection machinery differ with the use of feature selection and ensemble modeling methods?

5.2 Related Work

As discussed in Chapter 2, some researchers are modeling the dynamics of the meeting, exploiting dialogue models previously proposed for dialogue management. For example, Niekrasz et al. (2005) use the Issue-Based Information System (IBIS) model (Kunz and Ritte, 1970) to incorporate the history of dialogue moves into the Multi-Modal Discourse (MMD) ontology. Other researchers are modeling the content discussed in the meeting using the type of structures proposed in work on argumentation. For example, Rienks et al. (2005) have developed an argument diagramming scheme to visualize the relations (e.g., positive, negative, uncertain) between utterances (e.g., statement, open issue), and Marchand-Mailet (2003) propose a schema to model different argumentation acts (e.g., accept, request, reject) and their organization and synchronization. In these models, decisions are often seen as an outcome.

Automatically extracting the outcome of these argument models is still a challenging task. Researchers have just begun to make progress towards this goal. For example, Gatica-Perez et al. (2005) and Wrede and Shriberg (2003b) automatically identify the level of emotion in meeting spurts (e.g., group level of interest, hot spots). Other researchers have developed models for detecting agreement and disagreement in meetings, using models that combine lexical features with prosodic features (e.g., pause, duration, F0, speech rate) (Hillard et al., 2003) and structural information (e.g., the previous and following speaker) (Galley et al., 2004). More recently, Purver et al. (2006) have tackled the problem of detecting one type of decision, namely action items, which embody the transfer of group responsibility. However, no prior work has addressed the problem of automatically identifying decision-making units more generally in multi-party meetings. Moreover, no previous research has provided a quantitative account of the effects of different feature types on the task of automatic decision detection.

5.3 Methodology

The goal of work in this chapter is to develop models that can automatically detect decision-related conversations directly from the audio-visual recordings and to identify the feature combinations that are most effective for this task.

In Chapter 4, we empirically analyzed the features that are expected to be characteristic of decision DAs. Thus in this chapter, we focus is on how to computationally integrate the characteristic features to locate the decision-related DAs in meeting archives.

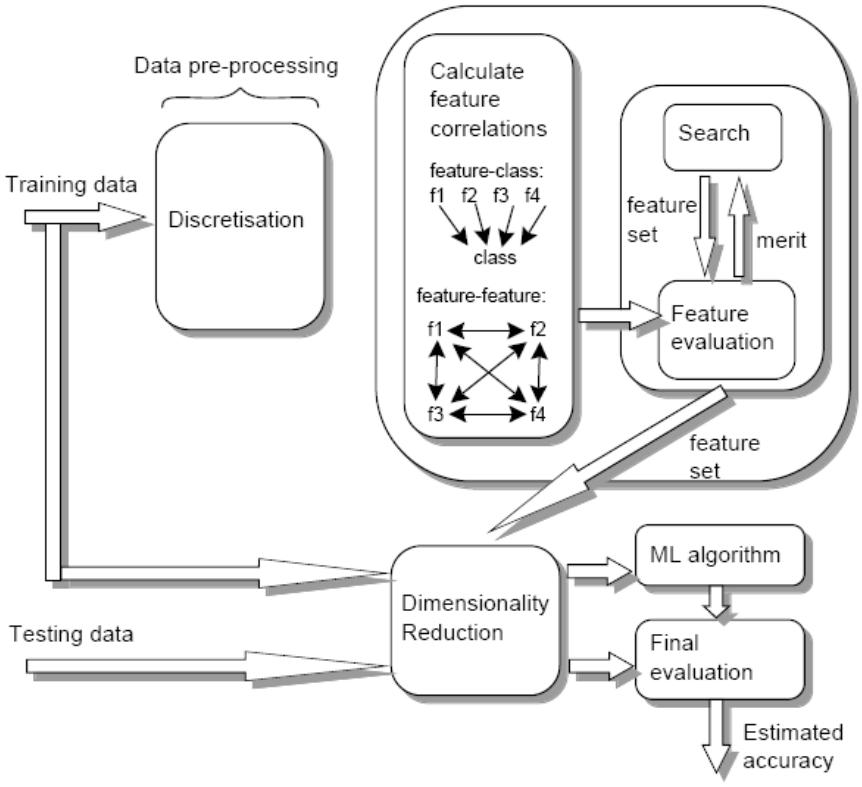

Figure 5.4: *Steps involved in the feature extraction and classification process (adapted from Hall (1998)).*

As discussed in Chapter 2, previous research on automatic meeting understanding has commonly utilized a classification framework, in which variants of generative and conditional models are computed directly from data. We also cast the decision detection task as that of classifying decision DAs, combining the wide range of decision characteristic features from multimodal inputs.

Figure 1.7 shows that how the decision detection task is positioned in the whole MDD system. The following Figure 5.4 further illusrates how the features extracted from the multimodal inputs are incorporated into the decision deteection task. Later in Section 5.3.4, we will describe more in details how the dimensionality reduction is done.

5.3.1 The Maximum Entropy approach

Many variants of classifiers have been used in spoken language understanding. In a pilot study, we experimented with decision trees, Support Vector Machines (SVM), and Maximum Entropy (MaxEnt) classifiers for detecting decisions.

Two biggest constraints in the requirement of a good classifer for the decision detection task are the capability of handling a large feature space (with 1K+ features) and imbalanced class distribution. The results of the pilot study reveal that these constraints have rendered the task as nontrivial. Among all, SVM and MaxEnt classifiers constantly outperform decision trees, yielding similar performance levels on detecting decision points. Both generalize relatively better in the presence of many features.

We first considered the appicability of SVM (Joachims, 1998) for the task. An SVM classifier learns an optimal threshold function $f(x) =< w,x > +b$, wherein x is the ith feature in the feature vector and w is the feature weight, to separates the training examples into two classes $\in \{+1(decision\,boundary), -1(non-decision\,boundary)\}$. However, the induced SVM classifiers tend to be not only more accurate on negative examples but also produce many false negatives which lead to low recall (Kubat et al., 1998). Yet in our pilot study the training of a better-performing SVM classifier tends to take much longer than training a MaxEnt one.

Then we considered the applicability of the MaxEnt classifiers Berger et al. (1996), which views the decision detection problem as a random process that produces an output class label y from a binary variable $Y : \{Yes, NO\}$, based on a contextual feature x, a member of a finite feature vector X. In our observation, MaxEnt models consistently perform well in this task, even with the presence of a large number of to-be-estimated parameters and an imbalanced class distribution. Previous work has shown MaxEnt classifier to be an effective tool for the sentence and discourse segmentation tasks (Liu et al., 2004; Christensen et al., 2005), which have constraints similar to our task.

There are several reasons for this. First, the MaxEnt classifier follows the principle of Occam's razor and shaves away features whose weights cannot be reliably estimated. Second, the classifier generalizes well to the imbalanced class distribution in the data, as it makes as few unnecessary independence assumptions as possible when fitting to the data. To simplify the presentation and to make the results more interpretable, we report only the results of the MaxEnt classifiers[1].

The principle of the MaxEnt classifier is to model all that is known, without making assumptions about what is unknown. That is, given all the data points, the MaxEnt classifier will find one model which is the most consistent with the data, but otherwise is as uniform as possible. In other words, it will incline to find a model which is the closest to uniform distribution such that the uncertainty ("surprise") is maximized.[2]

[1]The implementation we use in this thesis is the publicly available MaxEnt Modeling Toolkit provided by Zhang (2004). `http://homepages.inf.ed.ac.uk/s0450736/software/maxent/`.

[2]The interest of an uniform (highly uncertain) distribution emanates from the need for a distribution

Let $\tilde{P}(X,y)$, $Q(X,y)$ denote the empirical and the model distribution respectively: X represents the context in the form of a feature vector: $X = (x_1, x_2, ...x_d)$, with d equal to the length of the feature vector; y represents the target class selected from $Y \in \{+1, -1\}$. On the one hand, a MaxEnt model is constructed by adding features as constraints and adjusting weights to the features. Formally, the constraint in Equation 5.1 is used for model selection:

$$E_Q(f_i) = E_{\tilde{P}}(f_i), \forall f_i, \tag{5.1}$$

where $E_Q(f_i)$ and $E_{\tilde{P}}(f_i)$ denote the expectation with respect to the model distribution and the empirical distribution respectively, and $f_i(x,y)$ is the Boolean feature indicative function. Because MaxEnt models deal with binary features only, continuous features have to be discretized for MaxEnt. We applied a histogram binning approach, which divides the numerical value range into N intervals that contain an equal number of counts as specified in the histogram. The empirical expectation of $f_i(x,y)$ is computed using Equation 5.2:

$$E_{\tilde{P}}(f_i) = \sum_{x \in X, y \in Y} \tilde{P}(x,y) f_i(x,y). \tag{5.2}$$

$\tilde{P}(x,y)$ can be estimated as:

$$\tilde{P}(x,y) = \tilde{P}(x) P(y|x), \tag{5.3}$$

where $P(y|x)$ is the conditional distribution of (x,y) estimated by the MaxEnt model.

On the other hand, while complying to the above constraint, the MaxEnt model also attempts to obtain the optimal $Q(y|X)$ by maximizing the entropy ("expected surprise") – i.e., remaining as similar to the uniform distribution as possible.

$$argmax_\Lambda Pr(\tilde{P}|Q) = argmax_\Lambda \sum_{x \in X, y \in Y} \tilde{P}(x,y) log(P(y|x)). \tag{5.4}$$

Many iterative scaling and general purpose optimization algorithms have been proposed to estimate the parameters in Equation 5.4, e.g., Generalized Iterative Scaling (Darroch and Ratcli, 1972) and Improved Iterative Scaling (Pietra et al., 1997). For a detailed mathematical derivation, please refer to Pietra et al. (1997). The derivation arrives at an optimal model $\hat{Q}$ as follows:

in which we could not guess very well what a randomly drawn element will be. Such a distribution can ensure a fair situation with less bias introduced.

$$\hat{Q}(y|X) = \frac{1}{Z(X)} exp[\sum_{i=1}^{k} \lambda_i f_i(x,y)], \tag{5.5}$$

with the partition function[3] determined by $Z(X) = \sum_{y \in Y} exp(\sum_{i=1}^{k} \lambda_i f_i(x,y))$, with f_i represents the value of the *i*th feature in the model distribution.

In this study, the MaxEnt model is used for decision detection under the typical supervised learning scheme, that is, to train the classifier to maximize the conditional likelihood of the training data and then to use the trained model to predict whether an unseen spurt in the test set is a segment boundary or not.[4]

Furthermore, since it is often difficult to interpret a decision without knowing the current topic of discussion, we are also interested in detecting decision-related discussions at a coarser level of granularity: discourse segments. The task of automatic decision detection therefore is evaluated at these two levels of granularity: detecting decision DAs and detecting decision-related discourse segments.

In short, the automatic decision detection machinery consists of two components that locate decision-related information at the different levels of granularity: (1) a decision discourse dialogue act (DA) detector which identifies important DAs pertaining to the decisions made, and (2) a decision discourse segment (DS) detector which identifies the discourse segments in which interlocutors have reached one or more decisions.

The decision-related DS detector leverages the set of outputs (i.e., binary decisions) from the decision DA detector to classify whether an unseen discourse segment contains any decisions. The task of detecting decision-related segments thus can be viewed as that of recognizing decision DAs in a wider window (whose size depends on the length of the decision-related discourse segment the DA is located in), of which the size would be determined by an automatic algorithm discussed in Chapter 6.

5.3.2 Data

To provide training data for supervised training, in this study, we used a set of 50 scenario-driven meetings (approximately 37,400 DAs) from the AMI meeting corpus. These meeting recordings come with dialogue act annotations, extractive summaries and abstractive summaries. Decision-related DAs and discourse segments were also annotated (as noted in Section 3.2.3.1): the annotators determined for each dialogue

[3]The partition function is a special case of a normalizing constant in probability theory.

[4]In our experiments, the parameters of the MaxEnt model are optimized using Limited-Memory Variable Metrics (L-BFGS) (Malouf, 2002).

act in the extractive summary whether it is representative of some decision discussion and if so label it as a decision DA. Next, those discourse segments that contain one or more decision DAs were marked as decision DSs.

Although multiple annotations exist for some of the meetings, we used only one of the annotations, randomly chosen when multiple were available, as the ground truth data. On average, in the 50 meeting dataset, the annotators found eleven decision DAs and four decision DSs per meeting. Table 3.10 in Section 3.2.1.1 presents a break-down of the different types of decision-related segments. Overall, decision DAs account for 12.7% of the extractive summaries and 1.4% of the complete transcripts; decision-related segments account for 31.78% of the discourse segments in the dataset.

5.3.3 Feature extraction

In the corpus analysis of Chapter 4, we provided a qualitative account of the decision characteristics of the following features: lexical, prosodic, DA-related, topical, and subjective term features. The features of each conversation unit, i.e., dialogue act (DA), are represented as binary values in a feature vector, indicating whether each feature occurs within this unit or not. In this section, we introduce the set of potentially characteristic features that were used in the experiment and discuss how to extract them from the audio-visual recordings and the manual annotations. In addition, previous research has shown that some of the machine learning algorithms (e.g., instance based and naive Bayes classifiers) benefit from treating all features in a uniform fashion (Dougherty et al., 1995). Because MaxEnt models deal with binary features only, continuous features have to be discretized for MaxEnt. We applied a histogram binning approach, which divides the numerical value range into N intervals that contain an equal number of counts as specified in the histogram. This applies to any feature that contains numerical values.

Lexical Features

In Section 4.2, we showed that there are indeed lexical differences between general discussions and decision-related conversations. In prior work, lexical differences are often encoded by counting occurrences of cue phrases that have been empirically identified or reported in the literature. However, in a new domain, there is often no list of cue phrases available. To avoid the problem, in this study, we automatically identify

*B*ut um the feature that we considered for it not getting lost.
(but, um, the, feature, that, we, considered, for, it, not, getting, lost, ...) = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0...)
*U*m we so we do we've decided not to worry about that for now.
(um, we, so, we've, decided, not, to, worry, about, that, for, now, ...) = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0...)

Figure 5.5: *Lexical feature vectors for the example excerpt composed of decision DAs. Each vector is a binary variable representing whether or not each word has occurred in the excerpt.*

cue phrases from the corpus. Specifically, we compile a list of unigrams that were uttered in the decision DAs. Since the experiment in this chapter is conducted using a cross validation, in each fold, we compile such list from only those decision DAs in the training set of that fold.

Lexical feature vectors for the example meeting extract (as shown in Figure 5.5) were obtained by counting the occurrences of each of the decision-discriminative words in the list. If a word has occurred more than once, its feature value would be 1; otherwise, 0. In this study, we only count the occurrences of unigrams, as our exploratory study showed that including bigrams and trigrams does not improve over the accuracy of classifying decision DAs on the unigrams only.

Note that we did not apply stop word filtering in the generation of word vectors. Neither did we filter out the words that indicate disfluencies. Because in conversational speech, it is not easy to decide which words did not contain meanings. For example, Ferreira and Bailey (2004) have shown that including disfluencies (e.g., repeats, corrections, and editing terms such as "uh", "um") is essential to the computational models of parsing for language comprehension.

Prosodic Features

Functionally, prosodic features, i.e., energy, and fundamental frequency (F0), are indicative of segmentation and saliency. Prosodic features have been identified as indicative of different functional roles of discourse structure and in different types of recorded speech (Brown et al., 1980; Hirschberg and Nakatani, 1996; Wright, 1998; Tur et al., 2001; Hastie et al., 2002; Levow, 2004; Christensen et al., 2005). In Section 4.3, we identified prosodic patterns that distinguish decision-related conversations

from general discussions. In this study, we follow Shriberg and Stolcke's direct modeling approach to manifest prosodic features as duration, pause, speech rate, pitch contour, and energy level (Shriberg and Stolcke, 2001). Specifically, we utilize the cross-talking sound files, which were recorded by the microphone array in the center of the meeting table, in the AMI corpus. We did not use speech signals of individual close-talking microphones as they are distorted by heavy breathing, head-turning and cross-talk.

Type	Feature
Duration	Number of words spoken in current, previous and next DA
	Duration (in seconds) of current, previous and next DA
Pause	Amount of silence (in seconds) preceding a DA
	Amount of silence (in seconds) following a DA
Speech rate	Number of words spoken per second in current, previous and next DA
	Number of syllables per second in current, previous and next DA
Energy	Overall energy level
	Average energy level in the first, second, third, and fourth quarter of a DA
Pitch	Maximum and minimum F0, overall slope and variance
	Slope and variance at the first 100 and 200 ms and last 100 and 200 ms, at the first and second half, and at each quarter of the DA

Table 5.1: *An overview of prosodic features used in this study.*

DA-based Features

Our qualitative analysis in Section 4.4 indicated that contextual features specific to the AMI corpus, such as the speaker role (i.e., PM, ME, ID, UID) and meeting type (i.e., kick-off, conceptual design, functional design, detailed design), are characteristic of the decision DAs.

Analysis has demonstrated a difference in the dialogue act class, the reflexivity and the number of addressees – between the decision DAs and the non-decision DAs. For example, dialogue acts that belong to the classes, ***Inform***, ***Suggest***, ***Elicit-assessment***, and ***Elicit-inform*** are more likely to be decision-related.

Our analyses also found that the immediately preceding and following dialogue acts are important to identifying decision DAs. For example, we are more likely to see

a decision DA preceded by ***Stall*** and ***Fragment*** and followed by ***Fragment***. In contrast, there is a lower chance of seeing ***Suggest*** and elicit-type DAs (i.e., Elicit-information, Elicit-suggestion, Elicit-assessment) in the preceding and following decision DAs. Table 5.2 lists the contextual features used in this study.

DA Position in the meeting (in words, in seconds and in percentage)
Speaker role
Meeting type
Type of the current dialogue act
Type of the immediate preceding dialogue act
Type of the immediate following dialogue act
Reflexivity of the current dialogue act
Number of addressees of the current dialogue act

Table 5.2: *DA-based features in dialogue model.*

Topical Features

Thus far we have only considered features from within the analysis windows immediately preceding and following each potential boundary site. To explore models that take into account features of longer range dependencies, we further generated topic-specific language models and used the occurrence counts of words in the models as features to describe each potential boundary site. As described in Section 3.2.1.1, discourse segmentation and labels were also annotated in the AMI meeting corpus. Annotators had the freedom to mark a topic as subordinated (down to two levels) wherever appropriate. In this work, we flattened the structure into a hierarchy of two layers: top-level segments (TOPSEG) and all segments including subdiscourse segments (ALLSEG).

As noted in Section 3.2.1.1, because the AMI meetings are scenario-driven, we expected to find that most topics recur. Therefore, annotators were given a standard set of descriptions that can be used as labels for each identified discourse segment. In particular, the annotators explicitly identify those parts of the meeting that refer to the meeting process (e.g., opening, closing, agenda/equipment issues), or are simply irrelevant (e.g., chitchat). To capture the common characteristics of these off-topic discussion segments, we collapsed these segments into one category: functional segments (FUNC). The AMI scenario meetings take, on average, 30 minutes (around 800 DAs)

and contain an average of eight top-level discourse segments and seven sub-segments per meeting. (See Table 3.4 and Table 3.8 for a break-down of different types of segments.)

As reported in Section 3.2.1.3, we found that interlocutors are more likely to reach decisions when certain topics are brought up. In addition, we expected decision-making conversations to take place towards the end of a discourse segment. Therefore, in this study we included the following features: the label of the current discourse segment, the position of the DA in the discourse segment (measured in words, in seconds, and in %), the distance to the previous topic shift (both at the top-level and all levels of segments)(measured in seconds), the duration of the current discourse segment (both in the top-level and all levels of segments)(measured in seconds). Table 5.3 lists the set of topical features that are incorporated into our model to detect decision points.

Topic label
Position in a discourse segment (in words, in seconds, and in %)
Distance to the previous topic shift (both at the top-level and sub-topic level) (in seconds)
Duration of the current discourse segment (both at the top-level and sub-topic level) (in seconds)

Table 5.3: *Topical features in topic model.*

Subjective Term Features

In Section 4.5, we analyzed the distinctive use of subjective language in decision-related discussions. Our empirical analysis indicated that people do argue positively in the decision-related discussions, but often choose to express this in a relatively neutral way. For example, people use the word "*think*" quite often when discussing a particular decision – "*think*" is a word that has neutral sentiment polarity but positive arguing polarity. Another explanation is that there may exist a difference between the use of subjective expressions in the spoken and written language. Thus terms that are neutral in text may become subjective in meeting speech given the new context.

For our experiment, we used the MPQA Lexicon, a list of 8,221 subjective terms previously collected in text (Wilson et al., 2006). The list of terms are categorized

with respect to their subjective and sentiment properties.[5] Table 4.8 is a breakdown of the categories used in the subjective language analysis. For example, these terms are tagged with the degree of subjectivity, i.e., weak, strong, and the polarity of sentiment, i.e., positive, negative, neutral. In addition, these terms are also categorized with respect to their arguing polarity when used to express a belief or argument[6]. Table 4.9 further demonstrates the decision discriminability of the subjective terms of different sentiments and at different levels of polarity.

In this study, for each DA, a set of fifteen subjectivity categories are used (as shown in Table 5.4). The subjective features are computed by counting the number of times any of the subjective (or arguing) terms occurred in the dialogue act. Moreover, subjective terms of different categories are also counted as features, including the number of strong positive and weak positive terms. Arguing terms of different categories are also counted, including those with a positive arguing tendency and those with a neagative arguing tendency. As we have shown the importance of the neutral terms, we also counted those terms that have neutral sentiment and those where arguing tendency is neutral. Table 5.4 contains the list of the subjective term features we computed for each DA in the dataset used in our experiment.

Term Category	Term Features
Sentiment	positive, strong positive, weak positive, negative, strong negative, weak negative, polar, strong polar, weak polar, neutral, strong neutral, weak neutral
Arguing tendency	positive, negative, neutral

Table 5.4: *Subjective term features.*

5.3.4 Multimodal feature integration

Previous work has shown that combining multiple knowledge sources (e.g., words, audio-video recordings, speaker intention) is important to automatically identifying various phenomenon in human dialogues. For example, paralinguistic features (e.g., prosody and the amount of disfluency) have been applied to the detection of deceptive

[5]The properties were annotated in the list provided by Wilson et al. (2006).

[6]For example, the verb "think" is marked as positive, "deny" is negative, but the majority of words have a prior arguing polarity that is neutral, e.g., "wrath".

speech (Graciarena et al., 2006), utterance segmentation, disfluencies and pauses (Liu et al., 2004). Paralinguistic features have also been combined with features that indicate speaker intention (i.e., DA classes) to detect "hot spots", i.e., locations that exhibit a high level of affect in the voices of interlocutors (Wrede and Shriberg, 2003a,b). Similarly, lexical features, such as counts of cue words, have been used to detect learning attitudes of students using a tutoring system (Litman and Forbes-Riley, 2006) and to detect where speakers are agreeing with one another (Galley et al., 2004; Hillard et al., 2003).

The goal of our study is to develop a computational model to detect decisions from the audio-visual streams of human meeting dialogues. To make this computational model efficient, it is crucial to automatically identify a subset of features to keep out of the vast number of features extractable from the wide spectrum of knowledge sources. It is simply too computationally expensive to derive a model from all the available features. The high dimensionality of features will cause practical problems such as the enormous amount of data amassed from the large feature space, the slow learning process, and over-fitting of the classification models.

One way to avoid these problems is to reduce the feature space of individual knowledge sources to an optimal (or near optimal) subset. Feature selection is a long studied problem in the field of pattern recognition and machine learning (Allen, 1974; Langley, 1994; Hall, 1998). Some used the wrapper algorithm (Kohavi and John, 1996), which takes into account the bias introduced by the classifier and the task itself and finds useful, task-specific feature subsets by cross validating the performance of the classifier, using a two-phase scheme. It first prunes the features that do not degrade the performance when being left out in the model. Then, it performs a brute-force search over all possible subsets of features to identify those that maximize model performance.

Others have utilized filter algorithms (Das, 2001), which computationally characterize the potential merit of various feature subsets, with respect to some predetermined, classifier-independent criteria. For example, Hall (1998) proposed an algorithm to determine the relevant feature subset which has the strongest correlation with the target class. Yu and Liu (2003) proposed to further remove any redundant subset which completely correlates with another subset.

Complementary to the feature selection approaches that aim to select the potentially important feature subsets, a number of ensemble constructing algorithms have also been proposed to leverage libraries of models trained with selected feature subsets. These algorithms include meta-learning techniques, such as bagging (Opitz, 1999; Sut-

ton et al., 2005) and boosting (O'Sullivan et al., 2000).

The strategies used in this study to integrate features extracted from speech and video recordings are roughly divided into two categories with respect to the timing of feature fusion: early and late fusion. First, the early fusion strategy integrates multimodal information at the feature level by learning the correlation patterns of features across different knowledge sources. In the experiment of Section 5.8, we explore the use of different statistical and information theoretic discriminability measures, as well as different feature selection criteria, including:

- Maximum Decision Association (MA): Choose a subset of features on the basis of their discriminability of decision DAs according to a statistical measure, e.g., Pearson's Chi-Squared statistics (X2), or an information theoretic measure, e.g., information gain (IG).

- Maximum Decision Association, Minimum Feature Inter-correlation (MAMI): Choose a subset of features that have the highest discriminability of decision DAs and the lowest inter-correlations among themselves. E.g., Correlation-based feature selection (CFS) (Hall, 1998).

- Maximum Decision Association, Minimum Feature Redundancy (MAMR): Choose a subset of features that have the highest discriminability of decision DAs and remove those that are redundant. E.g., Fast Correlation-Based Filter (FCBF) (Yu and Liu, 2003).

The formulas listed below are modified from previous studies, which focus on measuring the correlation between features (e.g., Equation 5.6), to measure the feature correlation with regard to the class. Equation 5.7 to 5.9 is an example of using information gain in the measurement. (Previously in Chapter 4 we have reported the use of other measures such as LogLikelihood, Chi-Squared statistics and point-wise mutual information in our empirical studies.)

$$r = \frac{\sum(x_i - (\overline{x_i}))(y_i - (\overline{y_i}))}{\sqrt{\sum_i (x_i - (\overline{x_i}))^2}\sqrt{\sum_i (y_i - (\overline{y_i}))^2}} \tag{5.6}$$

$$H(X) = -\sum_i P(x_i) log_2(P(x_i)) \tag{5.7}$$

$$H(X|C) = -P(C)\sum_i P(x_i|C) log_2(P(x_i|C)) \tag{5.8}$$

$$IG(X|C) = H(X) - H(X|C) \tag{5.9}$$

if $IG(X|C) > IG(Z|C)$, the feature X is more correlated than the feature Z to the class C. To make this measure symmetrical between the feature and the class, the feature-class correltion is measured as symmetrical uncertainty. This measure removes the bias towards features of high values and normalizes the values to [0,1], with 0 indicating the feature is independent of the class and 1 indicating the feature can yield perfect prediction of the class.

$$SU(X,C) = 2[\frac{IG(X|Y)}{H(X)+H(Y)}] \tag{5.10}$$

Using the feature-class correlation measures, we can then select the feature subset either by searching through the feature subsets and identifying the set with the lowest feature-intercorrelation (e.g., Equation 5.11) or by adding non-redundant features one-by-one. The latter starts from traversing through the SU-ranked list of features $List_{SU} = (x_1, x_2, ...x_i, ...x_n)$ (wherein x_i is the ith feature) and including in the selected feature set $List_{BEST}$ only those that do not result in $SU_{BEST,x_i} >= SU_{BEST,C}$. The same process is iterated until all the features redundant to each of the features selected in $List_{BEST}$ are removed.

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}, \tag{5.11}$$

wherein M_s denotes the merit of feature set S (consisted of k features). $\overline{r_{cf}}$ and $\overline{r_{ff}}$ represent the average feature-class correlation and the average pair-wise feature correlation respectively. Table 5.5 presents a discriminative yet not-so-redundant feature subset selected from a meeting in the AMI corpus.

Second, the late fusion strategy integrates multimodal information at the knowledge source level through aggregating the classification results obtained with the individual sources.

In addition, as we expect these multimodal features to be correlated with one another, this study also explores how to select discriminant features in the early fusion stage. For example, this study examines the effect of the filtering criteria on the effectiveness and efficiency of the automatic decision detection machinery. Also, as we expect that the predictions yielded by the multimodal models to be complementary to each other, this study also investigates how to construct ensemble models from these models in the late fusion stage.

Type	IS1008c
Lex	display, advanced, contries, keep, slightly kinds, regarding, learn, decision, customisable par, consider, better, stick, phone acceptable, instruction, talked, useful, okay
Prosody	SHORT CURRENT ANSWER (DUR¡=2s) SHORT FOLLOWING RESPONSE (0.8s¡ FOLLOWING DA¡0.9s) RELATIVELY HIGH PITCH (6.9ms¡F0¡7.8ms) MID-RANGE SPEECH RATE (3.6¡Speech rate¡4.6 words) PITCH RISING IN Q1 and H1

Table 5.5: *Feature subsets that are the most discrimintive and the least redundant, selected with Symmetrical Uncertainty and FCBF search.*

Finally, as we are interested in examining the merits of multimodal feature combinations in this study, in the experiments 1-3 in Section 5.4 to 5.6, we quantify the impact of different knowledge sources wherever appropriate in the task of automatic decision detection. In the experiments 4-5 in Section 5.7 and Section 5.8, we then explore the use of early and late fusion strategies.

5.4 Experiment 1: Decision Detection from Extractive Summaries

Detecting decision DAs is the first step of automatic decision detection. For this purpose, we train MaxEnt models to classify decision DAs in the set of dialogue acts in the extracts, that is, those DAs that have been manually selected as extract-worthy. In this experiment, we want to focus on detecting decisions from extract-worthy DAs first in order to examine the effects of the different features on the task of decision detection in isolation. In Experiment 2, we train models to classify decision DAs directly on the entire transcripts, without requiring the extractive summaries.

To evaluate the performance of the models, we perform a 5-fold cross validation on the set of 50 meetings. In each fold, we train MaxEnt models from the feature combinations in the training set, wherein each of the extracted dialogue acts has been labelled as either POS or NEG. Then, the models are used to classify unseen instances in the test set as either POS or NEG. In Section 5.3.3, we described the four major types

of features used in this study: unigrams (LX1), prosodic (PROS), DA-related (DA), and topical (TOPIC) features. For comparison, we report the naive baseline obtained by training the models on the prosodic features alone, since the prosodic features can be generate automatically from pre-segmented recordings. The different combinations of features we used for training models can be divided into the following four groups: (A) using prosodic features alone (BASELINE); (B) using unigram, DA-based and topical features alone (LX1, DA, TOPIC); (C) using all available features except one of the four types of features (ALL-LX1, ALL-PROS, ALL-DA, ALL-TOPIC); and (D) using all available features (ALL).

Table 5.6 reports the performance on both the training (40 meetings) and the test set (10 meetings). Because previous work has shown that ambiguity exists in the assessment of the exact timing of decision DAs, context is needed to be seen by the system in order to disambiguate. Therefore, the results in Table 5.6 are also obtained using a lenient match measure, allowing a window of 20 seconds preceding and following a hypothesized decision DA for recognition. The right most three columns of the training set and test set results in Table 5.7 show the results of detecting decision-related discourse segments.

Each of these models will be evaluated with the accuracy of its predictions on decision-related units at the two levels: DAs and DSs. In the following experiments, we will report on the percentage of the predictions that match the actual decision-related units (*precision (P)*), the percentage of the decision-related units that have been correctly predicted (*recall (R)*), and the harmonic mean of the precision and recall(*F1*).

5.4.1 Decision-related dialogue act detection

Rows 1 and 2-4 in Table 5.6 report the performance of the BASELINE model and models in Group B, which are trained with a single type of feature. Lexical features are the most predictive features when used alone. We performed sign tests to determine whether there are statistical differences among the other models and the LX1 model. The baseline model is trained with prosodic features only. We do not use the randomly generated baseline, which makes class label predictions based on the probability of seeing a decision in a held-out development set. In out study, a dialogue act in the extractive summaries has only 12.7% chance of being a decision-related one. Under this setting, the random baseline will yield precision as 12.7%, recall as 12.7%, and F1 as 12.7%. The PROS baseline is harder to beat than the random baseline.

	TRAIN SET						TEST SET					
	Exact Match			Lenient Match			Exact Match			Lenient Match		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BASELINE	0.06	0.10	0.17	0.65	0.22	0.32	0.15	0.03	0.05	0.22	0.10	0.13
LX1	0.75	0.72	0.73	0.79	0.87	0.83	0.20	0.20	0.20	0.32	0.44	0.36
DA	0.52	0.01	0.02	0.62	0.02	0.04	0.22	0.01	0.02	0.24	0.01	0.03
TOPIC	0.60	0.09	0.16	0.73	0.13	0.22	0.22	0.05	0.07	0.35	0.08	0.13
ALL-LX1	0.72	0.38	0.50	0.81	0.60	0.68	0.28	0.18	0.22	0.46	0.35	0.40
ALL-PROS	0.84	0.70	0.76	0.89	0.86	0.87	0.31	0.24	0.27	0.45	0.39	0.41
ALL-DA	0.88	0.78	0.83	0.92	0.91	0.91	0.25	0.25	0.25	0.40	0.43	0.41
ALL-TOPIC	0.84	0.69	0.76	0.88	0.86	0.87	0.26	0.23	0.24	0.38	0.45	0.41
ALL	0.86	0.75	0.80	0.90	0.90	0.90	0.28	0.25	0.26	0.42	0.47	0.44

Table 5.6: *Effects of different knowledge sources on the accuracy (Precision (P), Recall (R), F1) of detecting decision DAs from extractive summaries. The right three columns of both the training set and test set results are obtained using a lenient match measure, allowing a window of 20 seconds preceding and following a hypothesized decision DA for recognition. Baseline is the prosodic feature-based model.*

Row 9 represents the performance of the ALL model, which combines the lexical, prosodic, DA and topical features. Comparing Row 9 with Rows 1-4 shows that the ALL model yields substantially better performance than the baseline on the task of detecting decision DAs.

To study the relative effect of the different feature types, Rows 5-8 in the table report the performance of models in Group C, which are trained with all available features except LX1, PROS, DA and TOPIC features, respectively. The amount of degradation in the harmonic accuracy (F1) of each of the models in relation to that of the ALL model indicates the contribution of the feature type that has been left out of the model. We performed sign tests to examine the differences among these models and the ALL model. We find that the *F1 score* (based on the *lenient match*) of the ALL model in the test set is better all of these models $(p < 0.001)$, indicating that all of these feature classes are important to the decision detection task to some extent.

A closer investigation of the *precision* and the *recall* of these models shows the following: First, removing the lexical and prosodic features would significantly degrade the *recall* whilst slightly increasing the *precision*. This shows that the decision-specific language and the prosodic models are essential to identify decision DAs. Second, re-

moving the DA-based and topical features would slightly decrease both the leniently measured *precision* and the *recall*.

5.4.2 Decision-related discourse segment detection

	Decision-related Discourse Segment					
	Training set			Test set		
Accuracy	Precision	Recall	F1	Precision	Recall	F1
BASELINE (PROS)	0.77	0.35	0.48	0.50	0.35	0.39
LX1	0.78	0.93	0.85	0.56	0.79	0.66
DA	0.82	0.03	0.06	0.40	0.05	0.09
TOPIC	0.85	0.16	0.27	0.58	0.16	0.24
ALL-LX1	0.88	0.68	0.76	0.65	0.56	0.59
ALL-PROS	0.91	0.89	0.90	0.62	0.62	0.62
ALL-DA	0.91	0.95	0.93	0.58	0.73	0.65
ALL-TOPIC	0.87	0.92	0.90	0.59	0.77	0.66
ALL	0.91	0.92	0.91	0.60	0.70	0.65

Table 5.7: *Effects of different combinations of features on detecting decision-related discourse segments from extractive summaries.*

As the task of detecting decision-related discourse segments can be viewed as a task of recognizing decision DAs in a wider window, the results in Table 5.7 are better than those reported in Table 5.6, achieving at best 91% harmonic accuracy (*F1*) in the training set and 65% in the test set. The model that combines all features (ALL) significantly outperforms all of the models that are trained with a single feature class, except LX1.

Rows 1-4 suggest that the lexical model (LX1), as compared to the baseline prosodic model (PROS) and the other models in Group (B) that are trained with a single feature class, is the most predictive in terms of harmonic accuracy (*F1*). Sign tests confirm the advantage of using LX1 ($p < 0.05$). Interestingly, the model that is trained with topical features alone (TOPIC) yields *precision* as good as using all of the features. This result stems from the fact that decisions are more likely to occur in certain types of discourse segments (c.f. Section 3.2.3.3). In turn, training models with topical features helps eliminate incorrect predictions of decision DAs in these types of discourse segments. However, the accuracy gain of the TOPIC model on detecting decision DAs in certain

types of discourse segments does not extend to all types of decision-related segments. This is shown by the significantly lower recall of the TOPIC model over the baseline ($p < 0.001$).

Finally, Rows 5-8 and Row 9 report the performance of the models in Group (C) and the model that is trained with all available features (ALL) on the task of detecting decision-related discourse segments. Calculating how much the harmonic accuracy of the models in Group C degrades from the ALL model shows that the most predictive features are the lexical features, followed by the prosodic features. Sign tests confirm that the ALL model outperforms the models that leave out lexical and prosodic features ($p < 0.05$). However, the ALL model does not outperform the model that leaves out DA-related and topical features due to the degradation of the recall.

5.4.3 Effects of combining lexical features with other feature classes

As the model that is trained with lexical features alone (LX1) yields harmonic accuracy as good as using all of the features, we are interested in knowing whether it is essential to combine lexical features with other types of features for the task of detecting decisions from extractive summaries. Table 5.8 further shows that combining prosodic, DA-related, and topical features with LX1 (LX1+PROS, LX1+DA, LX1+TOPIC) can improve the precision of the LX1 model but not the recall. This result stems from the fact that decision-characteristic words, such as content words, are also quite likely to appear in many other dialogue acts that are not directly related to decisions. Thus, combining other decision-characteristic features into the model helps eliminate incorrect predictions of decision DAs in these other non-decision DAs. However, this effect does not improve the recall of decision DSs. One possible conjecture is that most of the eliminated non-decision DA predictions are located in the same major discourse segments wherein interlocutors are likely to refer to the same terms.

In sum, we find that lexical models are indispensable for both the task of detecting decision DAs and discourse segments from extractive summaries. Also, the models that combine lexical, prosodic, contextual and topical features yield the best results on the task of detecting decision DAs, while models that combine lexical with any one of the other feature classes are sufficient for the task of detecting decision-related DSs.

	TRAIN SET						TEST SET					
Decision-Related	Dialogue Act			Dis. Segment			Dialogue Act			Dis. Segment		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BASELINE	0.65	0.22	0.32	0.77	0.35	0.48	0.22	0.10	0.13	0.50	0.35	0.39
LX1	0.79	0.87	0.83	0.78	0.93	0.85	0.32	0.44	0.36	0.56	0.79	0.66
LX1+PROS	0.85	0.87	0.86	0.84	0.93	0.88	0.37	0.47	0.41	0.59	0.76	0.67
LX1+DA	0.85	0.81	0.83	0.86	0.90	0.88	0.41	0.38	0.39	0.63	0.72	0.67
LX1+TOPIC	0.90	0.88	0.89	0.90	0.93	0.91	0.37	0.38	0.37	0.59	0.69	0.63
ALL	0.90	0.90	0.90	0.91	0.92	0.91	0.42	0.47	0.44	0.60	0.70	0.65

Table 5.8: *Effects of combining lexical and other features on detecting decision DAs and decision DSs from extractive summaries.*

5.5 Experiment 2: Detecting Decisions from Complete Recordings

As opposed to Experiment 1, which detects decision DAs from the set of DAs that have been identified as extract-worthy, in this experiment we trained models to detect decision DAs directly from entire transcripts of meetings. We expect this task to be much more challenging as the imbalance between positive and negative cases is even more prominent. The proportion of positive cases goes from 12.7% down to 1.4%. For comparison, we again use as a baseline the model based solely on prosodic features, which can be generated fully automatically. This is a harder to beat baseline than the randomly generated baseline in which a dialogue act has only a 1.4% chance of being a decision-related one.

Table 5.9 reports the performance on both the training (40 meetings) and the test set (10 meetings). Because previous work has shown that ambiguity exists in the assessment of the exact timing of decision DAs, in Table 5.9 we reported the results obtained by the lenient match measure. The right most three columns of the training set and test set results in Table 5.9 show the results of detecting decision-related discourse segments.

The results demonstrate that, compared to the PROS baseline and the semi-automatically generated LX1 model[7], models trained with all features (ALL), including lexical, prosodic, DA-related and topical features, yield notably better *precision* on the task

[7]Please note that the LX1 features used here are obtained on manual transcripts; so the lexical models can only be viewed as being trained semi-automatically.

	TRAIN SET						TEST SET					
Decision-Related	Dialogue Act			Dis. Segment			Dialogue Act			Dis. Segment		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BASELINE	0.73	0.00	0.01	0.73	0.01	0.02	0.32	0.01	0.01	0.51	0.03	0.06
LX1	0.40	0.60	0.48	0.55	0.81	0.66	0.26	0.48	0.33	0.48	0.81	0.60
ALL-LX1	0.80	0.13	0.22	0.90	0.17	0.28	0.44	0.09	0.14	0.63	0.21	0.31
ALL-PROS	0.86	0.57	0.68	0.90	0.66	0.76	0.37	0.21	0.27	0.61	0.49	0.53
ALL-DA	0.87	0.62	0.72	0.89	0.72	0.79	0.42	0.32	0.35	0.64	0.56	0.59
ALL-TOPIC	0.82	0.48	0.60	0.89	0.63	0.73	0.29	0.24	0.25	0.59	0.51	0.54
ALL	0.89	0.49	0.62	0.92	0.58	0.70	0.46	0.24	0.31	0.68	0.48	0.56

Table 5.9: *Effects of different combinations of features on detecting decision DAs and discourse segments from entire transcripts*

of decision-related discourse segment prediction, 92% on the training set and 68% on the test set. However, in the test set, the harmonic accuracy (56%) of the combined models is relatively worse than the LX1 model (60%), due to the substantially lower *recall* rate.

To study the relative effect of the different feature types, Rows 3-6 in the table report the performance of models in Group C, which are trained with all available features except LX1, PROS, DA and TOPIC, respectively. The amount of degradation in the harmonic accuracy (*F1*) of each of the models in relation to that of the ALL model indicates the contribution of the feature type that has been left out. For example, we find that the ALL model outperforms all except the model trained by leaving out DA-related features (ALL-DA). A closer investigation of the precision and recall of the ALL-DA model shows that including the DA-related features is detrimental to *recall* but beneficial for *precision*. This effect stems from the fact that decisions are more likely (1) to occur in certain types of dialogue acts, such as "Inform", "Suggest", "Elicit-Assess", and "Elicit-inform", and (2) to be preceded and followed by segmentation-type dialogue acts, such as "Stall" and "Fragment". Therefore, training models with DA-related features, such as the DA class of the current DA and its immediate context, helps eliminate incorrect predictions of decision DAs at the expense of *recall*.

In sum, the results suggest the following for the task of detecting decision points from entire transcripts: (1) Lexical features are the most predictive in terms of *har-*

	AllTran						Extract					
Decision-Related	Dialogue Act			DisṠegment			Dialogue Act			DisṠegment		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BASELINE	0.32	0.01	0.01	0.51	0.03	0.06	0.22	0.10	0.13	0.50	0.35	0.39
LX1	0.26	0.48	0.33	0.48	0.81	0.60	0.32	0.44	0.36	0.56	0.79	0.66
PROS+DA	0.28	0.02	0.03	0.49	0.08	0.12	0.37	0.22	0.27	0.64	0.47	0.52
PROS+TOPIC	0.39	0.06	0.11	0.67	0.13	0.22	0.41	0.31	0.34	0.58	0.54	0.55
DA+TOPIC	0.16	0.01	0.02	0.23	0.04	0.06	0.45	0.19	0.26	0.66	0.32	0.42
ALL-LX1	0.44	0.09	0.14	0.63	0.21	0.31	0.46	0.35	0.40	0.65	0.56	0.59
ALL	0.46	0.24	0.31	0.68	0.48	0.56	0.42	0.47	0.44	0.60	0.70	0.65

Table 5.10: *Effects of combining lexical and other features on detecting decision DAs and decision-related discourse segments. The first six columns are the results of operating the decision detection component on the whole recordings and the last six are on the part of recordings that have been previously selected as extractive summaries. ALL-LX1=PROS+DA+TOPIC.*

monic accuracy, despite low *precision*. Therefore, when harmonic accuracy is valued, the decision-specific language model is all we need to reach the best performance. (2) The non-lexical features, i.e. prosodic, DA-related and topical features, have positive impacts on *precision*. Therefore, when a more accurate decision detection component is needed, we should include the non-lexical features.

5.5.1 Effects of combining non-lexical feature classes

The results also suggest that, when the decision detection component is operated on entire transcripts as opposed to pre-selected extractive summaries, using the decision-specific language model to analyze what the speakers said is still the most effective way of identifying decision-related discussions. On the contrary, the importance of including non-lexical features (e.g., those extracted with the prosodic, dialogue, topic model) decreases. Therefore, a question naturally arises: What if the lexical features are not available? Can a reasonably performing model be trained with some of the non-lexical feature classes to detect decisions from the meeting recordings?

Table 5.10 shows the performance of the models trained with non-lexical feature combinations. None of these models performs comparably to those that are trained with lexical feature combinations. Only when all of the non-lexical feature classes (ALL-LX1) are combined can the system achieve comparable performance in *preci-*

sion, despite low *recall*. The low recall reveals that capturing the difference in expression styles (i.e., how the speakers said things) is not sufficient to the identification of decision-related discussions.

Capturing the difference in content (i.e., what the speakers said) is still essential. The decision DAs that can be captured by analyses of expression styles consist of only a small portion of all the decision DAs: 9% when only the audio-visual recordings are available and 35% when the extract-worthy part of recording is pre-specified. The discrepancy between the feature impact on decision detection from entire recordings and from extractive summaries suggests that the non-lexical models are not effective analysis tools for the task of distinguishing extract-worthy DAs.

When it comes to detecting decision-related discussion at the discourse segment level, the *precision* of the non-lexical models, 65%, is even better than the *precision* of the model that includes lexical models, 60%.

5.6 Experiment 3: Exploring Automatically Generated Features

5.6.1 Using automatically generated DA-class features

Our ultimate goal is to operate the decision detector in fully automatic fashion, and so we next evaluate the impact of automatically generated DA class features on the task of detecting decision DAs and discourse segments. To do so, we used the 5-class DA predictions (Auto-5DA) generated in (Dielmann and Renals, 2007b). We trained models which combine all available lexical, prosodic and topical features with the Auto-5DA features. The Auto-5DA model is evaluated against models that combine other features with the two annotated dialogue act class features: Manual-5DA (i.e., the manually annotated DA class) and Manual-15DA (i.e., the automatically generated DA class). The results reported here are obtained by operating the AMI Meeting Decision Detector on extractive summaries. In this way, we can focus on analyzing the impacts of the automatic DA features on the task of decision detection, rather than on that of extractive summarization.

Please note that because some of the test meetings we used in previous experiments are used as development set in (Dielmann and Renals, 2007b), the results reported here are obtained with a set of 50 meetings slightly different those used in previous experiments. Therefore a cross-table comparison of these results should be avoided.

	TRAIN SET						TEST SET					
Decision-Related	Dialogue Act			DisṠegment			Dialogue Act			DisṠegment		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Extract (Manual-15)	0.91	0.79	0.84	0.92	0.85	0.88	0.46	0.48	0.45	0.67	0.68	0.65
Extract (Manual-5)	0.88	0.88	0.88	0.87	0.92	0.89	0.45	0.56	0.49	0.64	0.79	0.70
Extract (Auto-5)	0.87	0.89	0.88	0.86	0.91	0.88	0.41	0.49	0.44	0.62	0.71	0.64
AllTran (Manual-15)	0.90	0.53	0.67	0.92	0.62	0.73	0.43	0.28	0.33	0.68	0.46	0.54
AllTran (Manual-5)	0.89	0.57	0.69	0.88	0.66	0.75	0.44	0.25	0.31	0.65	0.48	0.54
AllTran (Auto-5)	0.89	0.61	0.73	0.91	0.70	0.79	0.43	0.31	0.35	0.61	0.51	0.55

Table 5.11: *Effects of different versions of DA class features on detecting decision DAs and discourse segments. The first three rows (Extract) are the results obtained on extractive summaries. The last three rows (AllTran) are the results obtained on entire transcripts.*

Results in Table 5.11 show that our strategy that groups 15 DA classes into five major classes is beneficial to the models on the task of decision detection. On the task of detecting decision DAs from extractive summaries, it improves the recall of predicting decision-related discourse segments by 16%. Although replacing the manual 5-class DA features with the automatically generated version degrades the harmonic accuracy, the model trained with the 5 automatically predicted DA classes (Auto-5DA) still compares favorably with that trained with the 15 manually annotated DA classes (Manual-15DA).

However, when our system is operated on entire transcripts instead of extractive summaries, the advantage of the grouping strategy (from Manual 15-DA to Manual-5DA) does not exist. Neither is there any significant difference between the performance of Manual 5-DA and Auto-5DA. As in Experiment 5.5, we have observed that DA-related features are less predictive when predicting on entire transcripts. One conjecture is that DA-related features in general are not good at the dual task of disambiguating extract-worthy DAs and decision-related ones simultaneously.

5.6.2 Using automatically generated words

In Experiments 1 and 2, the LX1 features are extracted from manual transcripts; so the lexical models can only be viewed as being trained and tested in semi-automatic fashion. However, manual transcripts are costly to obtain, especially when the decision detection component is operated online or immediately following a meeting. There-

fore, if we want to develop a fully automatic component, we need to extract lexical features from the automatically recognized words, or use only non-lexical features. Experiment 3 shows that the latter approach is effective only when the component is operated on extractive summaries, due to the fact that only a portion of the decision DAs are distinguishable by the style in which they were expressed. Hence, we now examine the former approach.

As the previous experiments show that it is essential to combine lexical features with other features, it is important to know whether using the output of ASR will cause significant degradation to the performance of the decision detection models. In this study, we again use a set of 50 meetings and 5-fold cross validation. The meeting recordings come with manual and ASR transcripts obtained with the procedure described in Section 3.1.2. We compare the performance of the reference models, which are trained on human transcripts and tested on human transcripts, with that of the ASR models, which are trained on ASR transcripts and tested on ASR transcripts. In each fold, a decision-specific lexical model is trained and then used to find decision DAs and discourse segments from the unseen test meetings.

Columns 1-6 in Table 5.12 shows the results under the condition of operating on whole transcripts (AllTran) and Column 7-12 under the condition of operating on extractive summaries (Extract). These results indicate that when trying to identify decision DAs or segments in entire transcripts, there is no degradation incurred by using ASR transcripts.

However, the results obtained on extractive summaries show that using ASR transcripts impairs the performance of the lexical model significantly. However, using the ASR transcripts does not affect the performance of the combined model, in which the lexical features are integrated with all of the other available features. This suggests the features across knowledge sources are complementary. In those cases where the ASR errors would cause an incorrect prediction if used on their own, information conveyed in the prosodic, dialogue, or topic model will help to rectify it. Therefore, when the task is to detect decisions from extractive summaries and a high fidelity lexical model is not available, these other knowledge sources are indispensable to the development of a well-performing model. These results suggest an encouraging direction: that is, we can prevent the ASR recognition error-induced degradation in detection accuracy by including other non-lexical knowledge sources. However, as the state-of-the-art ASR system is not easily accessible and not all the meeting rooms are more noisy than those collected in the AMI corpus, the decision detection model would suffer from

	AllTran						Extract					
	Dialogue Act			Discourse Segment			Dialogue Act			Discourse Segment		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LX1 (REF)	0.26	0.48	0.33	0.48	0.81	0.60	0.32	0.44	0.36	0.56	0.79	0.66
LX1 (ASR)	0.26	0.47	0.33	0.47	0.80	0.59	0.28	0.47	0.35	0.50	0.76	0.60
ALL (REF)	0.42	0.28	0.33	0.65	0.45	0.53	0.46	0.48	0.45	0.67	0.68	0.65
ALL (ASR)	0.42	0.28	0.33	0.67	0.45	0.53	0.44	0.49	0.45	0.63	0.73	0.66

Table 5.12: *Effects of ASR words on detecting decision DAs and discourse segments. The first three columns (AllTran) are the results obtained on entire transcripts. The last three columns (Extract) are the results obtained on extractive summaries.*

high word error rates in real-life settings. On the one hand, this has further supported the importance of the use of non-lexical knowledge sources. On the other hand, we should also look for proxies of words that can be recognized with higher accuracy. In Section 6.5.2, we will report on the result of replacing ASR words with phonemes. We expect the same technique to be applicable to decision detection, but the actual implementation is beyond the scope of this thesis.

5.7 Experiment 4: Exploring the Use of Subjective Term Features

The empirical analysis in Section 4.5 suggested the decision-specific and subjective term-oriented language models are both characteristic of decision-related discussions. Whereas the decision-specific language models capture the content of the decision-related discussions in meetings, the subjective term-based models capture the expression style of the discussions. In Experiments 1-3, we have shown that learning decision-specific language models is an indispensable step in decision detection. Thus in this experiment, we examine the use of subjective term features in the development of a decision detection component.

The first question of interest in this study is whether including subjective term features helps train a well-performing MaxEnt model for decision detection to achieve the best performance, and which other features should be combined with the subjectivity-related features for such development. In this study, the subjective term features are calculated as the occurrence counts of subjective terms in the following 14 categories:

	AllTran						Extract					
Decision-Related	Dialogue Act			Discourse Segment			Dialogue Act			Discourse Segment		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SUBJ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LX1	0.26	0.48	0.33	0.48	0.81	0.60	0.32	0.44	0.36	0.56	0.79	0.66
LX1+SUBJ	0.28	0.41	0.33	0.51	0.74	0.60	0.36	0.39	0.37	0.57	0.69	0.62
ALL	0.40	0.22	0.28	0.68	0.46	0.54	0.43	0.47	0.44	0.61	0.70	0.65
ALL+SUBJ	0.46	0.24	0.31	0.68	0.48	0.56	0.41	0.49	0.44	0.60	0.72	0.65

Table 5.13: *Effects of subjective term features on detecting decision DAs and discourse segments. The first six columns (AllTran) are the results obtained on entire transcripts. The last six columns (Extract) are the results obtained on extractive summaries. ALL(+SUBJ)=LX1+PROS+DA+TOPIC(+SUBJ).*

strong subjective, weak subjective, strong positive, weak positive, strong negative, weak negative, strong neutral, weak neutral, strong positive arguing, weak positive arguing, strong negative arguing, weak negative arguing, strong neutral arguing, and weak neutral arguing terms.

The results in Table 5.13 show that using the 14 subjective term features alone does not yield useful decision detection model. Row 1 shows that none of the decision DAs can be predicted based on only the subjective cues. This can lead to the interpretation that the subjective cues are weak signals of decision-making conversations; their predictive power can only be revealed when they are combined with other cues. Combining the subjective term features with the decision-specific lexical features (SUBJ+LX1) can improve the *precision* of the LX1 model, which is trained with only the lexical features, but this is achieved at the cost of *recall*.

Another question of interest in this study is whether the subjective term features are needed when some of the other knowledge sources that are expected to capture the difference in speakers' expression styles are available, e.g., prosodic, dialogue and topic models. To answer this question, we also evaluated the performance of the combined ALL model, which is trained with all of the features (i.e., lexical, prosodic, dialogue, and topic features) used in the previous experiments, and that of the ALL+SUBJ model, which further includes the subjective term features. The results show that adding the subjective term features does not result in a significant performance gain. This experiment indicates that the information provided by the subjective term features

is redundant with the presence of other non-lexical features that are expected to capture the difference in expression styles.

5.8 Experiment 5: Multimodal Integration as Feature Selection

5.8.1 Comparing feature discriminability measures

Having established that it is possible to classify decision DAs and discourse segments given the lexical features (i.e., unigrams), we now assess the effect of feature discriminability measures on classification accuracy. Each measure assesses the feature discriminability of the unigrams for decision DAs so that the system can reduce the feature set by selecting the subset of unigrams that has the highest discriminability. In particular, this study employs four different lexical discriminability measures (which we have introduced in Section 4.2): Log Likelihood ratio (LL) and Chi-Squared statistics (X2), DICE coefficient (DICE), and Point-wise Mutual Information (PMI). While the latter two measure the association strength with information theoretic metrics, the former two measure with statistical metrics.

To compare the effectiveness of these measures, we adopt the following procedure. First, each of the four measures is applied to calculate the feature discriminability of all of the unigrams in the decision-specific language model, and the unigram features (i.e., words) are sorted with respect to their computed discriminability scores. Then, classification models are trained using the 25% most discriminating (Q1), the 25% mildly discriminating (Q2), the 25% mildly indiscriminating (Q3) and the 25% least discriminating (Q4) of these sorted unigram features. Finally, we examine the effect of features at different levels of lexical discriminability on the harmonic accuracy (F1) of detecting decision-related discourse segments from extractive summaries.

The intuition behind this procedure is that if any of these feature discriminability measures works well, the performance of the models trained using Q1 features selected by this measure ought to outperform the models trained using other subsets of less discriminative features. Table 5.14 suggests that Q1 features are always better predictors than Q2, Q3, and Q4 features, except for those selected by the PMI measure.

	Q1	Q2	Q3	Q4
LX1/LL	0.61	0.46	0.46	0.39
LX1/X2	0.60	0.50	0.51	0.40
LX1/DICE	0.61	0.41	0.28	0.31
LX1/PMI	0.49	0.51	0.44	0.36

Table 5.14: *Effects of feature discriminability measures on average classification accuracy (F1) of decision-related discourse segment models that are trained with unigram features (LX1). Q1, Q2, Q3 and Q4 features refer to unigram features selected at different levels of lexical discriminability in descending order according to the chosen discriminability measure.*

5.8.2 Comparing lexical and multimodal feature selection methods

With the effectiveness of feature discriminability measures confirmed, we next explore how to incorporate these measures into the feature selection process and find the optimal reduced feature set. Previous research has introduced a variety of feature selection methods, which take the selection criteria mentioned in the beginning of this subsection (i.e., MA, MAMI,) into account.

FS Criterion	Method	Average (Stddev) # of features in the model	Ratio to the original feature set (RATIO)
na	LX1	932.0 (± 14.7)	na
MA	LX1/X2	199.4 (± 20.6)	0.21
MA	LX1/IG	106.4 (± 9.9)	0.11
MAMI	LX1/CFS	95.6 (± 11.0)	0.10
MAMR	LX1/FCBF	102.6 (± 14.4)	0.11
na	ALL	1,506.0 (± 14.6)	na
MA	ALL/X2	210.0 (± 59.3)	0.14
MA	ALL/IG	269.0 (± 27.2)	0.18
MAMI	ALL/CFS	86.6 (± 6.5)	0.06
MAMR	ALL/FCBF	88.2 (± 15.9)	0.06

Table 5.15: *Effects of feature selection methods on the efficiency improvement of models. The last two columns show the size of the reduced feature subsets and the ratio of the subset size to the original set.*

Specifically, in this experiment, we examine the impact of the four feature selection

methods – Chi-squared statistics (X2), information gain (IG), correlation-based feature selection (CFS), and fast correlation-based filter (FCBF) – on both the efficiency and accuracy of the decision detection models. All four methods are tested on the LX1 model, which is trained with only the lexical features, and the ALL model, which is trained with the combination of lexical, prosodic, DA-based, topic, and subjective term features.

As one of the benefits of the different feature integration strategies is to reduce feature space, we first examine the ratio of the reduced feature set to the original feature space. Table 5.15 shows the potential impact of these feature selection methods on the creation of more efficient models. On average, the LX1 model is reduced to around 10%-21% of its original size, and the ALL model is reduced to around 6%-18%. Comparison of the feature subsets selected by the correlated-based feature selection (CFS) method and that selected by the Chi-squared statistic implies that the minimum feature intercorrelation criterion reduces the selected feature subset by half, from 21% to 10% for the LX1 model and from 14% to 6% for the ALL model. In contrast, comparison of the feature subsets selected by the FCBF method and those selected by the information gain method suggests that the minimum redundancy criterion produces efficiency improvements only on the ALL model but not the LX1 model.

Table 5.16 presents the impact of the four feature selection methods on the classification accuracy of decision-related discussions at both the dialogue act and discourse segment level. In addition to the precision, recall and harmonic F1 reported in Columns 4-9, in Column 3 the structural similarity of the model predictions is also reported. The **structural similarity** (*SSim*) is defined as the ratio of the number of model predictions to that of decision-related units in the reference data.

The first group refers to the results obtained with LX1 models. Row 1 refers to the original LX1 model. Row 2-5 refer to the reduced models yielded with the four feature selection methods. Results show that, among the feature selection methods that aim to maximize associations, the X2-reduced model performs better than the IG-reduced one in terms of the improvement in the precision of the original LX1 model on both the task of detecting decision DAs and discourse segments. Among the feature selection methods that do not only maximize associations but also apply extra filtering criteria, FCBF improves on the precision of models for both tasks, while CFS only improves on detecting decision DAs. However, as the reduced models consistently result in lower recall, the harmonic accuracy scores of these models are lower than the original LX1 model.

The second group refers to the results obtained with the ALL model that incorporates a wide range of multimodal features. Row 2-3 show that the X2 and IG feature selection method have improved on the precision of the original ALL model significantly on both the task of detecting decision DAs and discourse segments. Despite low recall, the harmonic accuracy at the discourse segment level remains similar to the original ALL model. Row 4-5 show that the CFS and FCBF method, which aim to minimize degree of intercorrelation and redundancy among features respectively, have degraded the precision at the DA level but improved it at the discourse segment level. This is possibly because the stricter criteria used in these two methods have resulted in the elimination of important features that are representative of some less common types of decisions; however, these criteria salso have preserved the predictive features for some most distinctive type of decision DAs, which are likely to appear at least once in every decision-related discourse segment. Hence the better performance in detection at the discourse segment level than at the DA level.

Overall, applying the feature selection methods to reduce the size of feature sets can improve the efficiency of models, but only achieve competitive performance when the reduced models are operated on the basis of all available multimodal features and used for detection at the discourse segment level.

5.9 Experiment 6: Multimodal Integration as Ensemble Modeling

In Experiment 5, we observed that using feature selection techniques in the early fusion stage tends to yield reduced models of high precision but low recall. As we expect each of these reduced models to capture the correlation patterns needed for identifying some distinctive types of decision DAs, in this experiment we present a method for constructing ensembles from libraries of these models in the late fusion stage. The problem of integrating multimodal information in the late fusion stage thus boils down to that of aggregating the classification results obtained with individual reduced models.

As a first attempt, Table 5.17 shows the results of constructing ensembles by voting. EMn refer to the number of votes (i.e., n) needed from the individual models for a positive final prediction. E.g., in EM3, a positive decision prediction is obtained only when at least three models predict that instance as decision-related. These results show

	Decision-Related		Dialogue Act			Discourse Segment		
FS Criteria	Method	SSim	P	R	F1	P	R	F1
n/a	LX1	1.01	0.32	0.44	0.36	0.56	0.79	0.66
MA	LX1/X2	0.35	0.34	0.17	0.22	0.59	0.41	0.48
MA	LX1/IG	0.40	0.32	0.20	0.24	0.56	0.47	0.50
MAMI	LX1/CFS	0.50	0.38	0.29	0.32	0.57	0.52	0.54
MAMR	LX1/FCBF	0.23	0.37	0.15	0.21	0.66	0.41	0.49
n/a	ALL	1.00	0.41	0.49	0.44	0.60	0.72	0.65
MA	ALL/X2	0.36	0.49	0.28	0.35	0.85	0.55	0.66
MA	ALL/IG	0.35	0.53	0.27	0.35	0.82	0.57	0.66
MAMI	ALL/CFS	0.51	0.32	0.22	0.25	0.75	0.57	0.64
MAMR	ALL/FCBF	0.36	0.31	0.15	0.18	0.79	0.51	0.62

Table 5.16: *Effects of feature selection methods on classification accuracy of decision detection models in extractive summaries. The first group are models that are trained with unigram features (LX1). The second group are models that are trained with the combination of lexical, prosodic, DA-based, topic, motion and subjective term features.*

that all of the ensemble LX1 models perform better than their source reduced models (c.f. Table 5.16) for detecting decision DAs, while the ensemble ALL models outperform the source models in this task as long as the needed votes are limited to under three. The vote constraint applies to the task of detecting decision-related segments for both the ensemble LX1 and ALL models.

We observe performance gains of *precision* in the results of all the ensembles of the reduced LX1 models and that of the reduced ALL models. The increasing *precision* rate from EM1 to EM4 matches the intuition that the more votes a target DA receives for its decision characteristics, the more likely this DA is a decision DA in the reference. However, as shown in Column 2 (*SSim*), raising the number of votes needed entails a decreased number of positive predictions and hence the lowered *recall* rate and *F1* harmonic accuracy.

In sum, the ensemble approach is favorable, especially to the LX1 models whose performance suffers greatly from early feature selection (c.f. Table 5.16).

Decision-Related		Dialogue Act			Discourse Segment		
Threshold	SSim	P	R	F1	P	R	F1
LX1	1.01	0.32	0.44	0.36	0.56	0.79	0.66
LX1/EM1	0.71	0.59	0.56	0.57	0.69	0.72	0.72
LX1/EM2	0.45	0.68	0.47	0.55	0.82	0.65	0.72
LX1/EM3	0.33	0.72	0.41	0.52	0.81	0.53	0.64
LX1/EM4	0.19	0.74	0.28	0.41	0.84	0.41	0.55
ALL	1.00	0.41	0.49	0.44	0.60	0.72	0.65
ALL/EM1	0.99	0.57	0.66	0.61	0.69	0.76	0.72
ALL/EM2	0.45	0.71	0.49	0.58	0.83	0.62	0.71
ALL/EM3	0.22	0.82	0.31	0.44	0.92	0.43	0.58
ALL/EM4	0.08	0.75	0.12	0.21	0.85	0.20	0.32

Table 5.17: *Effects of ensemble models on classification accuracy of decision DA and discourse segment detection models in extractive summaries.*

5.10 Discussion

People do speak and behave differently when expressing information about different aspects of the argumentation process. Our statistical and empirical analyses (as reported in Chapter 3) have shown that decision-related discussions do exhibit demonstrable differences in a wide range of features. In this chapter we explored the use of features from multiple knowledge sources (e.g., decision and subjective language model, prosody, dialogue, and topic model) for developing an automatic decision detection component for meeting speech.

Table 5.18 summarises the characteristics of decision-making discussions found in our empirical analysis reported in Section 4. We have anticipated these features to be predictive of the decision points). We have attempted to verify these hypothesis with the experiments reported in this chapter. Colmns 2-4 of Table 5.19 present the experimental results of how each type of the features fairs in the decision detection task when used alone. However, none of these feature types when used alone is powerful enough to train a well-performing decision detection model. The results suggest a modification to our hypothesis that takes the combination of indicaitive features into account.

Colmns 5-7 of Table 5.19 show how the combined model performs with each type

Feature	What to capture
Lexical	(1) There exists common expressions for decisions (e.g., we've decided). (2) Meeting participants are more likely to yield decisions when involved in the discussions of certain subjects (e.g., budget).
Prosodic	Meeting participants tend to express decisions in heightened speech (characterised by an increase in pitch slope followed by a decrease).
DA	(1) When expressing decision-critical information, meeting participants tend to express in the style of *informing, making suggestions,* or *eliciting others' assessment and information.* (2) Decision points are often preceded by a period when participants do not contribute more to the discussion but *stall or yield fragment (i.e. express nothing).* (3) The same happens aafter decisions are made
Topical	(1) Meeting participants are more likely to yield decisions when involved in the discussions of certain subjects (e.g., budget). (2) Decision-critical information occur either earlier on in a discussion (if given by the agenda) or near the end (when consensus is reached).
Subjective	Speakers tend to use more words that are neutral but with positive arguing power in decision-related discussions

Table 5.18: *Decision-discriminative feature types and the systematic differences they capture.*

Feature	Standalone performance			Performance when removed		
	P	R	F1	P	R	F1
Lexical	–	–	–	+	–	-
Prosodic	–	—	–	+	–	-
DA	–	—	—	-	-	-
Topical	–	—	–	-	-	-
Subjective	—	—	—	+	-	o
Lexical+ Prosodic	-	o	-	n/a	n/a	n/a

Table 5.19: *Decision-discriminative feature types, how each type fairs in the decision detection task when used alone, and how the combined model (which integrates all the feature types) performs with each of these feature types removed. The mark in each cell indicates the level of difference between the leniently matched accuracy of this model and that of the combined model for detecting decision-related DAs from extractive summaries: + (better than the combined model); o (no difference); - (worse); – (at least 10% worse); — (the model does not work at all with accuracy under 0.1).*

of these features removed. The leave-one-out performance levels demonstrate that the lexical, prosodic, and subjective features come as double-bladed: including them in the combined model decreases recall in exchange for precision. In the follow-up experiments reported in this chapter, we have also found that when not all of the features are available, the most indispensible combination of feature types are that of the lexical and prosodic features. That is to say in order to know when the meeting participants are talking about decision-critical information, we need to observe not only what speakers says but also how they express themselves. (However, finding the indicative combinations of individual features, e.g., if a speaker is talking about budgets in high pitch, is beyond the scope of this thesis.)

5.10.1 Can decision-related conversations be detected automatically?

As the first step towards answering the question, we presented the meeting decision detector (MDD), a system which performs automatic decision detection in meeting speech and provides visual aids for users who wish to review decisions. The task

of predicting decision-related discussions were evaluated at two levels of granularity: "dialogue act" (DA) and "discourse segment (DS)". A MaxEnt approach was used to classify each potential dialogue act or discourse segment into the class of being a decision-related one or not.

We first examined how our computational models perform when they are used to classify decision DAs from extractive summaries, i.e., a manually selected set of dialogue acts that are representative of what has happened in a meeting. To overcome the problem of imbalanced class distribution – on average, only 12% of the dialogue acts in the extractive summary are decision DAs – we leveraged a variety of knowledge sources (e.g., decision-specific language, prosody, DA-related context, topic model). From these knowledge sources, we extracted a wide range of features are expected to capture the difference in content (i.e., what the speakers said) and expression style (i.e., how the speakers said it) to train the classifier.

For identifying decisions at the dialogue act level, the results suggest that simply by training a decision-specific language model and using it as a basis to construct lexical models, a MaxEnt classifier can yield reasonable performance on detecting decision points, significantly outperforming the baseline (i.e., the automatically generated prosodic model). In addition, incorporating lexical features with features extracted from other knowledge sources, such as the prosodic, dialogue and topic model, yields combined models that significantly improve both the precision and the recall of the lexical model in the task of recognizing decision points from extractive summaries at the dialogue act level.

At the discourse segment level, the task of detecting decisions is essentially a task of recognizing decision DAs in a wider window. The results show that the combined model does not outperform the lexical model in this task. The additional features impair the recall rate at the discourse segment level, even though they have been proven to improve recall at the DA level. Counterintuitive as this result may seem, further analysis reveals that certain types of discourse segments, by nature, are less likely to have decisions reached in them. For example, those segments wherein meeting participants are discussing agenda items. As a result, the additional knowledge sources, specifically the topic model, are inclined to eliminate correct predictions of decision DAs made by the lexical model in these types of discourse segments, yielding a lower recall rate. This is why although these non-lexical knowledge sources improved the precision of the model, they do not yield a distinctive harmonic performance gain in the task of recognizing decisions from extractive summaries at the discourse segment

level.

In sum, the results suggest the following for the development of a well-performing automatic decision detector: (1) Lexical features are essential to maintaining the recall and therefore must be included in the model. (2) The additional knowledge sources, i.e., prosodic, dialogue and topic model, are important to improving the precision of the lexical model. Therefore, when a more accurate decision detection component is needed, these knowledge sources should be incorporated into a combined model.

As the empirical study in Chapter 4 has shown that people do speak differently in decision-related discussions, we also examined whether incorporating features that characterize the use of subjective language can further improve the performance of the decision detection models. The results show that further combining subjective features can improve the precision of the detection model despite lower recall.

5.10.2 Can the decision detection component be operated fully automatically?

One drawback of our approach is that it requires human intervention, such as manual transcripts, manual extractive summaries, manually annotated DA segmentations and labels, and other types of meeting-specific features (e.g., speaker role). It is therefore essential to examine whether a well-performing decision detection model could be developed without the aid of human intervention. First, we examined whether the MDD system is robust to other noise introduced by the automatically generated versions of other features. We also used the automatically generated version of the DA class features, i.e., automatic DA classifications (as reported in (Dielmann and Renals, 2007b)). The insignificant negative impact we observed shows that it is possible to incorporate the automatic DA class features in the model directly.

Next, to determine whether it is necessary to operate the decision detection model on the manual extractive summaries, we evaluated our computational models on the complete meeting recordings. Detecting decisions from entire recordings is expected to be a more difficult task, since only 1.4% of the dialogue acts in complete transcripts are positive cases, and this task essentially involves classifying decision-related and extract-worthy dialogue acts simultaneously. The evaluation showed that using solely the lexical model can achieve the best performance, 60% *harmonic accuracy* for detecting decision DSs. This is 10% worse than the results obtained on the extractive summaries, but still much better than the chance level given the imbalanced class dis-

tribution consists of only 1.4% positive cases.

Further including the non-lexical knowledge sources yields a combined model that performs substantially better in terms of precision, achieving 46% and 68% on the task of detecting decision DAs and discourse segments respectively. In sum, these results suggest that it is possible to develop an automatic decision detection model to be operated on entire recordings directly. When precision is more important, we should use the combined model; When recall or harmonic accuracy is primary consideration, we should use the lexical model. Furthermore, the results also suggest that the automatic model is robust to the noise introduced by the automatically generated features.

5.10.3 How to integrate multiple knowledge sources effectively?

Having established a platform to evaluate decision detection models and identified the effective knowledge sources for decision classification accuracy, the next question is whether it is possible to automatically reduce the large number of features needed for developing a well-performing model and reach similar or even better performance than the original model.

In this study, we explored both the statistical and information theoretic measures and confirmed their effectiveness in discriminating decision-related discussions. I have then attempted to accommodate these measures of lexical discriminability in the feature selection algorithms so that the multimodal features extracted from the widely varying knowledge sources can be integrated in the early fusion stage. With the reduced models developed, we finally combined the predictions made by these reduced models to integrate the multimodal information at the late fusion stage.

Explicitly, for early fusion we used three feature selection criteria, which endeavor to maximize the association between the features and the decision DAs, minimize the inter-correlations among the features, and filter out the redundant features. For late fusion we experimented with a simple voting technique to construct an ensemble model from the reduced models.

Overall, the combined model was reduced to around 6% to 18% by the early fusion methods, and its performance is as good as the original model. That implies the reduced models can be five to 15 times more efficient than the original model, achieving the same level of effectiveness. However, this finding does not hold true when the early fusion methods are used to generate reduced lexical models. Although the lexical model can also be reduced to around 10% to 21% of its original size, the re-

duced model degrades the performance of the original model by 10% to 15%, despite the merit of the discriminability measures on characterizing the association strength between the lexical features in the model and the decision DAs.

In sum, our results show that when the early fusion methods are applied to select features from multiple knowledge sources, they can yield reduced models that are more efficient than the original model. Further constructing an ensemble model from the reduced models can increase the harmonic accuracy of the decision detection models.

5.11 Summary and Limitations

In this chapter, we have cast the task of decision detection as that of classifying decision DAs and discourse segments from the extractive summaries. The hypothesized decision DA information is then incorporated into the display of a meeting browser to enable efficient browsing and search of meeting decisions. In particular, we have explored with five ways to model the multimodal characteristics of decision-related conversations: a decision-specific language model, prosody model, dialogue act-based model, topic model, and subjective language model. While the decision-specific language model captures the differences in what the speakers said during a meeting (i.e., "content"), the other knowledge sources capture the differences in how the speakers said it (i.e., "expression style"). A series of experiments have been performed to provide a quantitative account of the merits of these various models in use.

The comparison of the feature merits has suggested the following for the development of a well-performing automatic decision detector: (1) Lexical features are essential to good performance on recall, and therefore, it is necessary to incorporate the decision-specific lexical features in the model. (2) The additional knowledge sources, i.e., prosody, dialogue act, topic and the subjective language model, are important for improving the precision of the lexical model. When a model that issues fewer false positives is preferred, it is necessary to combine the lexical model with some of these knowledge sources, most effectively the dialogue act-based model, into a combined model. (3) When the trade-off between precision and recall is of more concern, combining the lexical model with more than one of the non-lexical models is necessary for achieving the best harmonic accuracy. Among these non-lexical models, the prosodic model is the most important to be included, followed by the dialogue model, the topic model, and the subjective language model.

As the development of the model requires considerable human intervention, we

further examined whether it is possible to develop a decision detection system that can be operated in a fully automatic fashion. As a first step, we attempted to relax the requirement on having to operate on extractive summaries. We compared the results across the condition of operating on extractive summaries and that of operating on entire transcripts (c.f. Table 5.6 and 5.9). We found that the lexical features are important no matter which condition is in operation, albeit with a greater positive impact when used in a model that is operating on complete recordings. This is because, in addition to the task of classifying decision DAs, the development of a model in complete recordings also involves the task of classifying extract-worthy DAs, a task whose success relies heavily on the lexical features.

The impact of other, non-lexical, knowledge sources is less prominent in the task of detecting decisions on entire recordings. No combination of the non-lexical features can yield a good performing model in terms of the harmonic accuracy. But these combinations do yield models of high precision (c.f. Table 5.10). When combined with the lexical model, all of the non-lexical knowledge sources improve the precision, but seriously degrade the recall. This suggests that, for this more difficult task the non-lexical features are a more reliable discriminator but only for a small portion among the various types of decision DAs. Therefore, only when the goal is to develop a more precise model will the non-lexical features add value to the combined model. The features ranked in descending order of their impact on precision are the topic, prosodic, dialogue, and subjective language model.

Having established a platform for evaluating merits of the features and identified the effective knowledge sources for the development of automatic decision detection models, we then explored whether early or late fusion techniques can further improve the efficiency and effectiveness of the developed models. we found that although the reduced models produced by the early fusion methods are 5 to 15 times more efficient than the original model, they are only, at best, as effective as the original model. However, if we apply the late fusion method to construct an ensemble model from these reduced models, we can improve the harmonic accuracy of the original model by an additional 10%, yielding 61% and 72% for the task of detecting decision DAs and segments respectively.

Despite these encouraging results, our decision classification approach has some inherent limitations, which stem from the fact that this approach essentially views the decision summarization task as that of compiling an automatically selected set of DAs as an extractive summary of the meeting decisions. First, the unconnected dialogue

acts in the excerpt can resulted in semantic gaps that would require contextual information to bridge. Second, anaphora and unexpected topic shifts between these extracted DAs also require contextual information to resolve. Finally, although it is our intuition that the decision DA extracts will assist users in finding and absorbing information in the meeting archives, this assumption has yet to be tested with human subjects.

To address the first liability and provide the needed contextual information, in the next chapter (Chapter 6), we train computational models to find contextual information that is needed for interpreting the identified decisions. In particular, we explore methods to find discourse segment boundaries, for example, where a new discourse segment or discussion of this decision was initiated. The identified segments are then used to automatically indicate the topic of the current decision-related discussion.

Finally, to address the third limitation, in Chapter 7, we report on an extrinsic evaluation using the decision debriefing task and provide a quantitative account of the utility of displaying decision DA information (as exemplified in Figure 1.6) to users performing this task common in our daily organizations life.

(1) A: but um the feature that we considered for it not getting lost.
(2) B: Right. Well
(3) B: we're talking about that a little bit
(4) B: when we got that email
(5) B: and we think that each of these are so distinctive, that it it's not just like another piece of technology around your house.
(6) B: It's gonna be somewhere that it can be seen.
(7) A: Mm-hmm.
(8) B: So we're we're not thinking that it's gonna be as critical to have the loss
(9) D: But if it's like under covers or like in a couch you still can't see it.
...
(10) A: Okay , that's a fair evaluation.
(11) A: Um we so we do we've decided not to worry about that for now.

Figure 5.6: *Example decision-making discussion*

Chapter 6

Meeting Discourse Segmentation: Determining Relevant Contexts

6.1 Introduction

Annotating implicit semantics to enhance browsing and searching of recorded speaker interaction speech poses new challenges to the field of natural dialogue understanding. Studies in human information retrieval show that structuring retrieved information in a hierarchical display helps users find information more efficiently (Dumais et al., 2001). Usability research also elucidates the benefit of an intuitive navigation interface which is tagged with labels that reflect the users' mental organization (Resnick and Sanchez, 2004; Rosenfeld and Morville, 2002). Although these studies were done in the context of web browsing, the findings suggest the usefulness of displaying contents along with its structural information.

In the context of spoken language understanding, structural information has also been shown to be useful in many spoken language understanding tasks, including anaphora resolution (Grosz and Sidner, 1986), information retrieval (e.g., as input for the TREC Spoken Document Retrieval task), and summarization (Zechner and Waibel, 2000). Moreover, it also lends support to the development of dialogue systems by improving dialogue act and speech recognition (Hastie et al., 2002).

Similar benefits have been observed in the context of meeting information retrieval. First, annotating transcripts with structural information (e.g., topics) enables users to browse and find information from multimedia archives more efficiently (Kominek and Kazman, 1997; Banerjee et al., 2005). Second, discourse segmentation makes up for the lack of explicit orthographic cues (e.g., topic and paragraph breaks) in speech.

One critical problem to overcome is thus how to automatically divide unstructured multi-party interaction into a number of locally coherent segments. The application needs are two-fold. First, the recognized discourse segment structure forms a quasi-summary and serves as an overview of what has transpired in a meeting. The goal is to provide end users the right level of detail to interpret what has happened in a meeting. Let's return to the scenario we used in Chapter 1. Suppose you are an industrial designer who has missed a meeting and wanted to review the design team's discussion about the target user group. If the system can provide a discourse segment structure as shown in Figure 6.1, you can then efficiently locate the information you are looking for (in this case, the segment about "target user group") from the list of segments. As evidenced in (Banerjee et al., 2005), discourse information does enable users to browse and find information from a meeting archive more efficiently. Moreover, when a recorded meeting has to be displayed on a mobile device, the recognized discourse segments can be used to construct an easy-to-grasp, thumb-nail view of the meeting. In short, discourse segmentation recognition has great potentials to enhance the current user interaction scheme of browsing and search.

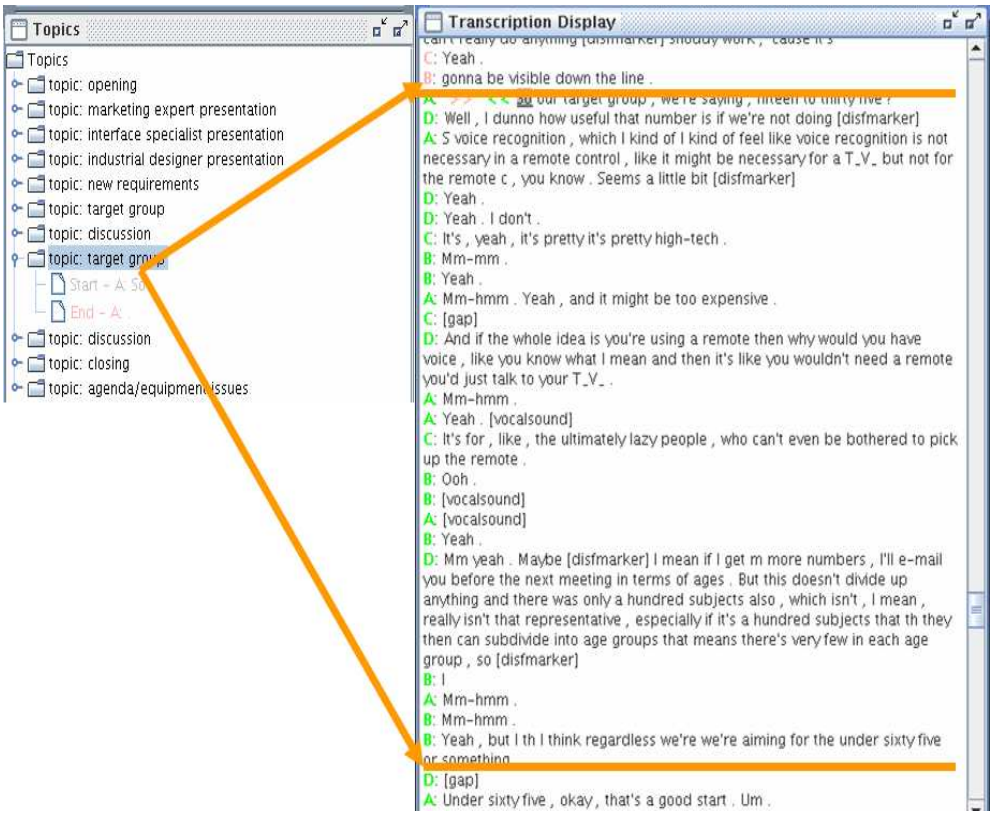

Figure 6.1: *Example of discourse segmentation in a produce design meeting.*

Second, discourse segment recognition benefits the development of other downstream meeting understanding applications. These applications include anaphora resolution (Grosz and Sidner, 1986), information retrieval (e.g., as input for the TREC Spoken Document Retrieval (SDR) task), summarization (Zechner and Waibel, 2000), and question answering. Ostendorf et al. (2008) have pointed out the need for optimizing segmentation for these end tasks, rather than independently.

As in Chapter 4 and Chapter 5, we have shown the benefits of integrating multiple knowledge sources for meeting dialogue understanding. Following this line of research, in this chapter we explore whether automatic machinery can be developed to integrate multiple knowledge sources for segmenting recorded speech. Specifically, we cast the task of automatic discourse segmentation in two distinctively different approaches: first as an unsupervised semantic similarity-based clustering task, and then as a supervised feature-based classification task. The former approach follows the LCSeg approach (Galley et al., 2003), first put forth in Hearst (1997), to find optimal segmentation by locating lexical changes over meeting speech, while the latter approach applies the feature-based classification approach similar to that used by the decision detection component developed in Chapter 5.

We first compare the two segmentation approaches and examine the impact of lexical information and multiparty interaction cues, e.g., overlap, pause, speaker activity change, which have been used in previous work on meeting segmentation. Next, we examine the use of other knowledge sources that have not been previously studied systematically in previous work, but which we expect to be good predictors of discourse segments in dialogue. These include prosodic and motion cues. In Experiment 1, we provide a quantitative account of the effectiveness of the two segmentation approaches and the merit of the different knowledge sources.

Because tasks as diverse as browsing, on the one hand, and summarization, on the other, require different levels of granularity of segmentation, we also examine the impact of these different knowledge sources on the accuracy of segmenting three different types of discourse segments: (1) hypothesizing where major topic changes occur (TOPSEG), (2) hypothesizing where more subtle nested topic shifts occur (ALLSEG), and (3) hypothesizing where off-topic functional discussions occur (FUNC). While previous work focuses on finding only the major topic segments (TOPSEG) (Garofolo et al., 2000; Galley et al., 2003), in this chapter we also explore useful features and models for the tasks of recovering sub-level and functional discourse segments.

Last but not least, as ultimately we hope to operate the segmentation component

in an online scenario and in a fully automatic fashion, we also investigate the impact of using only automatically extractable information on the accuracy of the two segmentation approaches. To determine the impact on the supervised feature-based classification approach, we construct models that are developed with the manual features and that with automatic ones. We then examine whether the performance of the model would be degraded by replacing the manual features with their automatically generated versions, such as replacing manual transcription with ASR transcripts.

To determine the impact on the unsupervised clustering approach, we examine the effectiveness of approaches that can operate directly on audio sources. First, we examine the performance change in the segmentation approach caused by replacing manual transcripts with ASR transcripts. Next, to avoid the requirement of high fidelity transcription, we also examine the performance change caused by replacing word-based transcripts with phonetic transcripts, which can be obtained directly from audio inputs in near real time. In Section 6.5.2, we describe how the speaker activity-enhanced phonetic representations are processed and how the changes in repetitions of phonemes and that of speaker activities are located. In Experiment 6, we compare our audio-based system against those systems that operate on meeting transcriptions.

In sum, in this chapter we address the following research questions:

1. (1) How to adapt the methods previously developed in text and broadcast news segmentation to segment meetings, integrating the various potentially predictive knowledge sources?

2. (2) What are the effective knowledge sources serving for finding discourse segments of different types?

3. (3) How can we adapt the offline segmenters to be operated online or immediately following end of a meeting?

To answer the first question, we look at whether lexical information is sufficient for the development of a well-performing segmenter, and whether integrating the potentially characteristic multimodal and multiparty interaction features can improve the performance of the segmenters. To answer the second question, we perform a study to investigate the difference between the effective knowledge sources for recovering the two-layer discourse structure and the off-topic functional discourse segments and what features in these sources are most useful for recognition. The third question is answered by examining whether the use of ASR transcripts seriously degrades per-

formance and whether an online automatic discourse segmenter can be developed to operate directly over the audio sources of meetings.

In examining the impact of the different segmentation approaches, knowledge sources and discourse types on the multiparty dialogue segmentation task, a series of experiments are first performed on the ICSI meetings. In Experiment 1, we compare the impact of lexical information and multiparty interaction cues on the accuracy of discourse segmentation at two different levels of granularity: TOPSEG and ALLSEG. In Experiment 2, we select the discourse boundary-signalling cue phrases using the statistical lexical discriminability measures that have been proven useful in Chapter 4 and evaluate the impact on segmentation accuracy. To adapt the off-line model to be used in online scenarios, in Experiment 3, we evaluate the impact of operating on ASR transcription.

To determine if our results generalize to other types of meetings and to experiment with additional features not available in the ICSI corpus (e.g., motion features), another series of experiments are performed using the scenario-driven meetings available in the AMI meeting corpus. Many more types of knowledge sources, widely ranging from motion and prosody to dialogue context, are then scrutinized for their merits to be included. Experiment 4 provides a quantitative account of the effect of using different knowledge sources on segmentation accuracy. To study the applicability of the developed segmentation component in the online scenario, in Experiment 5 and Experiment 6, we report the impact of the ASR words and that of using the unsupervised segmentation approach directly over the audio sources of meetings.

6.2 Related Work

In Section 2.5, we identified three categories of discourse segmentation approaches that have been proposed to recover simple discourse structures from lengthy dialogues (like those in meetings). A comparison of the three categories of models is presented in Table 2.3. Because one of the goals of this thesis is to look for approaches that can be operated online, the computationally expensive sequence coding approach is not viable in this research. In this section, we discuss the remaining approaches, which are relevant to our problem.

6.2.1 Unsupervised semantic clustering

The problem of automatic discourse segmentation is often considered as similar to the problem of text segmentation. Therefore, researchers have adopted time series analysis approaches, which were previously developed to segment topics in text (Kozima, 1993; Hearst, 1997; Reynar, 1998) and in read speech (e.g., broadcast news) (Ponte and Croft, 1997; Allan et al., 1998; Trieschnigg and Kraaij, 2005). For example, lexical similarity-based algorithms, such as LCSeg (Galley et al., 2003) or its word frequency-based predecessor TextTiling (Hearst, 1997), capture topic shifts by modeling the similarity of word repetition in adjacent windows. Some systems have already achieved good results for predicting topic boundaries when trained and tested on human transcriptions. For example, Stokes et al. (2004) report an error rate of 0.25 on segmenting broadcast news stories using the unsupervised similarity-based approach.

However, recent work has shown that LCSeg is less successful in identifying "agenda-based conversation segments" (e.g., *presentation*, *group discussion*) that are typically signalled by differences in group activity (Hsueh and Moore, 2006). This is not surprising since LCSeg considers only lexical similarity in terms of cohesion.

6.2.2 Audio-video features beyond words

Many studies in conversation segmentation have studied how to combine features automatically extractable from knowledge sources other than words, such as cross-speaker linking information (e.g., adjacency pairs (Zechner and Waibel, 2000)) and participant behaviors (e.g., note taking cues (Banerjee and Rudnicky, 2007)). In fact, previous work has already shown that training a segmentation model with speaker interaction features (e.g., overlap rate, pause, and speaker change) (Galley et al., 2003) can outperform LCSeg on the task of segmenting meetings.

In many other fields of research, a variety of features have been identified as indicative of segment boundaries in different types of recorded speech. For example, Brown et al. (1980) have shown that a discourse segment often starts with relatively high pitched sounds and ends with sounds of pitch within a more compressed range. Passonneau and Litman (1993) identified that topic shifts often occur after a pause of relatively long duration. Other prosodic cues (e.g., pitch contour, energy) have been studied for their correlation with story segments in read speech (Tur et al., 2001; Levow, 2004; Christensen et al., 2005) and with theory-based discourse segments in spontaneous speech (e.g., direction-given monologue) (Hirschberg and Nakatani, 1996). In

addition, head and hand/forearm movements are used to detect group-action based segments (McCowan et al., 2005; Al-Hames et al., 2005).

However, many other features that we expect to signal segment boundaries have not been studied for the merits of including them to build predictive models. For instance, speaker intention (i.e., dialogue act types) and speaker interaction context (e.g., number of addresses). In addition, although these features are expected to be complementary to one another, few of the previous studies have examined the correlation among features.

6.2.3 Supervised feature-based classification

Many different techniques have been proposed to construct models from features that have been determined to correlate with audio-video segment boundaries, such as cue phrases, part-of-speech tags, co-reference relations (Gavalda et al., 1997; Beeferman et al., 1999). These models were then used to classify whether each of the potential boundary sites is a segment boundary or not. For example, Brown et al. (1996), van Mulbregt et al. (1999), and Blei and Moreno (2001) use topic language models and variants of the hidden Markov model (HMM) to identify topic segments in multimedia documents.

However, discourse segmentation of multiparty dialogue seems to be a considerably harder task. Recordings of multiparty dialogue lack the distinct segmentation cues commonly found in text-like transcripts (e.g., headings, paragraph breaks, and other orthographic cues) or news story segmentation (e.g., the distinction between anchor and interview segments). Galley et al. (2003) reported an error rate (P_k) of 0.319 for the task of predicting major discourse segments in meetings.[1] (For a more detailed overview of previous work in discourse segmentation, please refer to Chapter 2.)

6.2.4 Moving from off-line to on-line scenario

As one of the goals in this research is to study how to operate the discourse segmentation component in an online or near online scenario, in this chapter we also discuss the feasibility of fully automating the segmentation approaches that have been previously proposed, including both the unsupervised semantic similarity-based and the supervised feature-based classification ones.

Among the unsupervised segmentation approaches it is observed that the success of previous work in this area depends largely on the quality of text-based transcripts.

[1] For the definition of P_k and W_d, please refer to Section 6.3.2.

Although it is natural for the researchers who studied semantic similarity-based approaches to assume the availability of manual transcripts or, at least, ASR outputs, high fidelity transcription is difficult to obtain in a timely manner.

The search for approaches that do not require the existence of high fidelity transcription has led us to the field of spoken language understanding, in which researchers have proposed a few more time series analysis approaches that can work directly on audio sources, without having to transcribe the signals into words first. In particular, we notice studies on locating changes over the acoustic units. For example, Malioutov et al. (2007) used an unsupervised vocabulary acquisition technique (Park and Glass, 2006) to derive sub-lexical units (i.e. those corresponding to high frequency words and phrases). Using this approach, inter-utterance similarity could be measured in ways similar to text segmentation (Utiyama and Isahara, 2001; Choi et al., 2001). However, it is uncertain whether the vocabulary acquisition algorithm that was developed to work on monologues (e.g., lectures) would be robust to processing recordings of multiparty meeting dialogues.

Then, we turn to the studies of supervised approaches and find most of the automatic segmentation models in prior work were also developed for off-line scenarios. Only some of these studies reported on performance degradation caused by replacing manual transcription with ASR transcripts. For example, past research on broadcast news story segmentation using ASR transcription has shown performance degradation from 5% to 38% using different evaluation metrics (van Mulbregt et al., 1999; Shriberg et al., 2000; Blei and Moreno, 2001).

Compared to broadcast news and two-party dialogue, multi-party dialogues typically exhibit a considerably higher word error rate (WER) (Morgan et al., 2003). However, no prior work has examined the extent to which incorrectly recognized words would impair the unsupervised segmentation approach and the extraction of conversation-based discourse cues and other features used in the feature-based classification models. We therefore will provide a quantitative account of the impact of the ASR errors in the following experiments of these two approaches.

Among the few studies that attempted to construct segmentation models with automatically generated features, most of them detected changes in speaker activity from the audio inputs (Renals and Ellis, 2003; Galley et al., 2003). As noted in this review, there are many more features that are expected to be predictive of segment boundaries. Therefore, in the following experiments, we will also assess whether the benefit of including these automatic, non-lexical features compensates for the errors introduced in

the automatic feature extraction process. In addition, as we believe that previous work has not fully exerted the potential of audio-based approaches, we will also experiment with unsupervised acoustic similarity-based approaches, in which acoustic units are monitored and significant changes are detected in near real-time.

6.3 Methodology

6.3.1 Data

This study aims to explore approaches that integrate multimodal information to segment multiparty conversations at two levels of granularity.

As our goal is to identify multimodal cues of discourse segment boundaries in face-to-face conversation, we use both the ICSI meeting corpus and the AMI meeting corpus, which contain the recordings of spontaneous multiparty meeting dialogues. (For more detailed description of these corpora, please refer to Chapter 3.)

To evaluate the performance of various features on discourse segmentation in meeting speech, we first segment a recorded meeting into minimal units, which can vary from sentence chunks to blocks of sentences. In this study, we use *spurts*, that is, consecutive speech with no pause longer than 0.5 seconds, as minimal units, as they are automatically determinable. We take each of the spurts which the annotators have chosen as the first spurt of a segment to be the boundary sites.

Then, to study how to segment meetings at the coarse and fine level of granularity, we characterize a dialogue as a sequence of segments that may be further divided into sub-segments. we take the discourse segmentation annotations in the corpus and flatten the sub-segment structure and consider only two levels of segmentation: top-level segments (TOPSEG) and all segments including sub-level segments (ALLSEG). In Section 3.2.1.1, we described the procedure for annotating both the ICSI and the AMI corpus with this type of hierarchical segmentation scheme. Additionally, as we wish to study whether some of the segmentation models are more beneficial than others at predicting off-topic discourse segments, we label all off-topic segments that are related to opening, closing, chitchat, and agenda/equipment discussions in the AMI corpus as functional segments (FUNC) and evaluate the accuracy of these models on detecting the FUNC segments.

6.3.2 Evaluation metrics

An automatic segmentation system involves two error types: *false negative* (i.e., failures to detect a boundary site) and *false positives* (i.e., false alarms of a boundary site that does not exist). When there is a trade-off between error types, a single performance number is often inadequate to reflect the overall capabilities of a system. Nevertheless, various metrics that aim to use a single number have been proposed in the field of text segmentation to evaluate the performance of segmentation models. One typical example is to evaluate the performance in terms of cross-class accuracy – i.e., the number of correct predictions out of all predictions (including both the positive and the negative ones). Previous work has shown that when class distributions display a high level of entropy, i.e., $P(c_i \mid T) \approx P(c_j \mid T), i \neq j$ for any two classes c and training data T, cross-class accuracy is an acceptable measure of quality for a detection system.

However, discourse segmentation is a typically class-imbalanced task. The number of linguistic units on which segmentation is based, e.g., utterances, by far exceeds the number of actual segments. Consequently, optimizing a classifier for accuracy would automatically favor a majority classifier that labels all sentences as not initiating a new segment. Optimization for the classical notions of accuracy scores that measure only the correct predictions of a target class – *recall* (i.e., the number of correct predictions out of all boundary sites in the ground truth) and *precision* (i.e., the number of correct predictions out of all predicted boundary sites) – would not work well here either. For instance, a discourse segmenter that always predicts a segment boundary close to but not exactly corresponding to the ground truth boundary sites would produce zero recall and precision, while the performance of this segmenter should be considered as good.

6.3.2.1 P_k and W_d

In response to these problems, P_k and W_d were designed to overcome the limitations inherent in the use of precision and recall for discourse segmentation. Beeferman et al. (1999) defined the P_k measure as the probability that a randomly drawn pair of utterances (k utterances apart) are incorrectly predicted as coming from the same segment.

Pevzner and Hearst (2002) analyzed several weaknesses of the P_k measure, including the undesirability of P_k in overpenalizing false negatives.[2] To remedy these

[2]In the simplest cases where only one false negative exists, the penalty for a false negative is k, whereas in the simplest case where only one false positive exists, the penalty for a false positive is only

weaknesses, they proposed an adapted metric WindowDiff (W_d). As formalized in Equation 6.1, W_d is computed as the probability that the number of hypothesized and ground truth segment boundaries in a given window frame are different.

$$W_d = \sum_{i=1}^{N-k} \frac{[|r(i,k) - h(i,k)| > 0]}{N-k}, \tag{6.1}$$

where $r(i,k)$ is the number of boundaries between position i and $i+k$ in the ground truth data, and $h(i,k)$ is that in the predictions. However, as false positives and false negatives are normalized by the same factor $N-k$ in W_d, W_d has been reported to penalize false positives more than false negativees in Georgescul et al. (2006).[3]

In this chapter, we provide an aggregated account of both P_k (Beeferman et al., 1999) and W_d (Pevzner and Hearst, 2002). Note that P_k and W_d values indicate segmentation error rates; therefore, the lower the P_k or W_d value is, the closer the hypothesized segmentation is to ground truth, with 0 signaling perfect segmentation.

6.3.2.2 SDis: Structural distance

However, evaluating with the P_k or W_d scores can be tricky, since the scores are not normalized for the number of segments a model hypothesizes. Therefore, we also report on the structural similarity (*SDis*) of the segmenter in evaluation. Let HYP be the number of the system-hypothesized segments, and REF be the number of the ground truth segments. As shown in Equation 6.2, *SDis* is defined as the ratio of the difference between REF and HYP to REF, with directional information specified in the sign. A negative SDis score indicates that the model in test predicts fewer decision-related DAs than the reference ones.

$$SDis = \frac{HYP - REF}{REF} \tag{6.2}$$

The closer to zero, the more similar the hypothesized segment structure is to ground truth. The SDis figure then tells us whether the target segmenter exhibits any irregular under-segmentation or hyper-segmentation behavior that should be considered along with its P_k or W_d score. We also consider scenarios in which the number of ground

j ($j < k$), the distance from where it occurs to the nearest boundary site.

[3] Since the number of boundary sites in the ground truth only consists of a small portion of all possible boundary sites, the chance of seeing a false negative is expected to be lower than that of seeing a false positive.

truth segments is unknown to the segmenter, and thus it is necessary to study how accurate the system is in predicting the right number of segments in a meeting.

6.3.2.3 Recall: Finding functional segments

In some cases we are interested in further analyzing the accuracy of the segmenters on different types of segments, specifically the off-topic functional segments. In these cases, we also report on the segment-type specific recall. For example, for measuring how accurate a segmenter's prediction is on recovering the functional segments, the recall is defined as the proportion of the hypothesized boundaries that correspond to at least one of the ground-truth functional segment boundaries.[4]

6.3.2.4 DET curve

Finally, a segmentation system has many operating points, and is best represented by a performance curve. Therefore, in addition to reporting the segmentation error rate, the recall rate of off-topic functional segments and the structural distance, we also plot the DET (Detection Error Trade-off) Curve to see how the segmentation models work at different operating points.

6.4 ICSI Meeting Segmentation

In this chapter, we first address the challenge of whether we can segment a multiparty dialogue recording directly over the meeting transcripts without training. To do this, we apply a time series analysis approach, LCSeg (Galley et al., 2003), which hypothesizes that a major topic shift is likely to occur where strong term repetitions start and end. The algorithm works with two adjacent analysis windows, each of a fixed size which is empirically determined. LCSeg calculates a lexical cohesion score by computing the cosine similarity at the transition between the two windows. Low similarity indicates low lexical cohesion, and a sharp change in lexical cohesion score indicates a high probability of an actual segment boundary. The principal difference between LCSeg and TextTiling (Hearst, 1997) is that LCSeg measures similarity in terms of

[4] We do not report on precision as the segmenters tested in this study are not tailored to find functional segments.

lexical chains (i.e., term repetitions), whereas TextTiling computes similarity using word counts.

Figure 6.2 exemplifies how drastic changes in lexical cohesion scores correspond with the discourse segment boundaries chosen by the annotators. Although not all of the drastic change points have a correspoonding manual segment boundary, there is only one segment boundaries that does not come with a drastic change within a window of 3 minutes. This phenomena suggests that there exists differnet types of segment boundaries – some of them are related to topic shifts while some others are not. The TextTiling approach is predictive of the topic-based discourse segmentation.

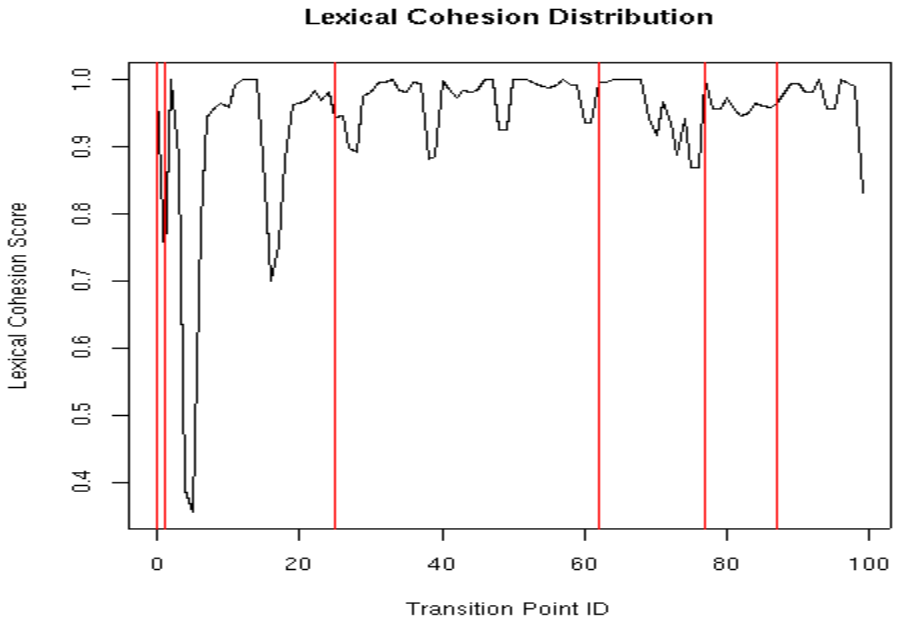

Figure 6.2: *Change of lexical cohesion scores over an example ICSI meeting. The red lines represent segment boundaries specified by human annotators.*

We then cast the task of discourse segmentation as that of classifying each potential segment boundary site into the category of segment boundary (POS) or non-segment boundary (NEG). Here, the potential segment boundary sites are determined by the boundary site of spurts, i.e., consecutive speech with no pause longer than 0.5 seconds. Under the typical supervised learning scheme, the task is to train a classifier to maximize the conditional likelihood over the training data and then to use the trained model to predict whether an unseen spurt in the test set is a segment boundary or not.

For this study, we first trained decision trees (c4.5) to learn the best indicators of segment boundaries. We constructed feature vectors for all spurts with the follow-

ing features: (1) lexical cohesion features: the raw lexical cohesion score (LCV) and probability of topic shift (LCP) indicated by the sharpness of change in lexical cohesion score, (2) cue word features: the occurrence count of cue words in an analysis window of 5 seconds preceding and following the potential boundary (COL-CUE)[5], and (3) speaker interaction features: including the amount of speaker activity change (measured as the distance between the two probability distributions of words spoken by each speaker) within 5 seconds preceding and following each potential boundary (SP_k), the amount of overlapping speech within 30 seconds following each potential boundary (OVR), and the amount of silence between speaker turns within 30 seconds preceding each potential boundary (SIL). [6]

While the unsupervised LCSeg approach uses only lexical information, the supervised machine learning approach further integrates information from multiparty interaction cues, such as overlap, pause, and speaker activity change. Galley et al. (2003) showed that the supervised learning approach outperforms LCSeg in the task of automatic discourse segmentation at the top level. Our objective here is thus to determine whether integrating multiparty interaction features also improves automatic discourse segmentation at the finer granularity, as well as when ASR transcriptions are used.

To compare to prior work, we performed a 25-fold leave-one-out cross validation on the set of 25 ICSI meetings that were used in Galley et al. (2003). In each evaluation, we trained the automatic segmentation models for two tasks: predicting all segment boundaries (ALLSEG) and predicting only top-level segment boundaries (TOPSEG). In order to be able to compare our results directly with previous work, we first report our results using the standard error rate metrics of P_k and W_d.

Baseline

To compute a baseline, we follow Kan (2003) and Hearst (1997) in using Monte Carlo simulated segments. For the corpus used as training data in the experiments, the probability of a potential segment boundary being an actual one is approximately 2.2% for all sub-level segments, and 0.69% for top-level discourse segments. Therefore, the Monte Carlo simulation algorithm predicts that a speaker turn is a segment boundary with these probabilities for the two different segmentation tasks. we executed the al-

[5]Cue words reported in Galley et al. (2003) include “ok”, “okay”, “and”, “anyway”, “alright”, “but”, and “so”.

[6]The window sizes are selected based on those reported to perform best in Galley et al. (2003) for segmenting ICSI meeting transcripts into major segments.

gorithm 10,000 times on each meeting and averaged the scores to form the baseline for our experiments.

Topline

For the 24 meetings that were used in training, we have top-level topic boundaries annotated by coders at Columbia University (COL) and in our lab at Edinburgh (EDI). We take the majority opinion on each segment boundary from the COL annotators as ground truth segments. For the EDI annotations of top-level discourse segments, where multiple annotations exist, we choose one randomly. The Topline is then computed as the P_k score comparing the COL majority annotation to the EDI annotation.

6.4.1 Experiment 1: Off-line ICSI discourse segmentation from human transcripts

In order to facilitate meeting browsing and question-answering, we believe it is useful to include sub-level discourse boundaries in order to narrow in more accurately on the part of the meeting that contains the information the user needs. Therefore, we perform experiments aimed at analyzing how the LCSeg and machine learning approach behave in predicting segment boundaries at the two different levels of granularity.

All of the results are reported on the test set. Table 6.1 shows the performance of LCSeg, which predicts based solely on lexical information, and the combined model (CM), which is trained using C4.5 with a combination of the lexical cohesion, cue phrase and speaker interaction features as discussed in Section 6.4. For the task of predicting top-level discourse boundaries from human transcripts, CM outperforms LCSeg. LCSeg tends to over-predict on the top-level, resulting in a higher false alarm rate. However, for the task of predicting sub-level discourse shifts, LCSeg alone is considerably better than CM.

Next, we wish to determine which features in the combined model are most effective for predicting topic segments at the two levels of granularity. Table 6.2 gives the average P_k for all 25 meetings in the test set, using the features described in Section 6.4. we group the features into four classes: (1) lexical cohesion-based features (LF): including lexical cohesion value (LCV) and estimated posterior probability (LCP); (2) interaction features (IF): the amount of overlapping speech (OVR), the amount of silence between speaker segments (GAP), similarity of speaker activity (ACT); (3) cue

Error Rate		Transcript		ASR	
Models		P_k	W_d	P_k	W_d
LCSeg	ALLSEG	0.323	0.382	0.329	0.371
	TOPSEG	0.365	0.466	0.380	0.482
CM	ALLSEG	0.369	0.387	0.382	n/a
(C4.5)	TOPSEG	0.284	0.295	0.284	n/a

Table 6.1: *Performance comparison of the two probabilistic segmentation approaches.*

phrase feature (CUE); and (4) all available features (ALL). For comparison we also report the baseline (see Section 6.4) generated by the Monte Carlo algorithm (MC-B). All of the models using one or more features from these classes outperform the baseline model. A one-way ANOVA revealed this reliable effect on the top-level segmentation task ($F(7,192) = 17.46$, $p < 0.01$) as well as on the task of segmenting all segments including subdiscourse segments ($F(7,192) = 5.862$, $p < 0.01$).

TRANSCRIPT	Error Rate(P_k)	
Feature set	ALLSEG	TOPSEG
MC-B	0.466	0.484
LF (LCV+LCP)	0.381	0.299
IF (ACT+OVR+GAP)	0.389	0.301
IF+CUE	0.389	0.301
LF+ACT	0.387	0.301
LF+OVR	0.386	0.295
LF+GAP	0.385	0.299
LF+IF	0.381	0.296
LF+CUE	0.375	0.292
ALL (LF+IF+CUE)	0.369	0.284

Table 6.2: Effects of feature combinations for predicting topic boundaries from human transcripts. MC-B is the randomly generated baseline.

As shown in Table 6.2, the best performing model for predicting top-level segments (TOPSEG) is the one using all of the features (ALL). This is not surprising, because these were the features that Galley et al. (2003) found to be most effective for predict-

ing top-level segment boundaries in their combined model. Looking at the results in more detail, we see that when we begin with LF features alone and add other features one by one, the only model (other than ALL) that achieves significant improvement ($p < 0.05$) over LF is LF+CUE, the model that combines lexical cohesion features with cue phrases.[7]

When we look at the results for predicting sub-level discourse boundaries, we again see that the best performing model is the one using all features (ALL). Models using lexical-cohesion features alone (LF) and lexical cohesion features with cue phrases (LF+CUE) both yield significantly better results than using interaction features (IF) alone ($p < 0.01$), or using them with cue phrase features (IF+CUE) ($p < 0.01$). Again, none of the interaction features used in combination with LF significantly improves performance. Indeed, adding speaker activity change (LF+ACT) degrades the performance ($p < 0.05$).

Therefore, we conclude that for predicting both top-level and sub-level discourse boundaries from human transcriptions, the most important features are the lexical cohesion based features (LF), followed by cue phrases (CUE), with interaction features (IF) contributing to improved performance only when used in combination with LF and CUE.

However, a closer look at the P_k scores in Table 6.2, adds further evidence to our hypothesis that predicting sub-level discourse segments may be a different task from predicting top-level discourse segments. Sub-level segment shifts occur more frequently, and often without clear speaker interaction cues. This is suggested by the fact that absolute performance on sub-level discourse prediction degrades when any of the interaction features are combined with the lexical cohesion features. In contrast, the interaction features slightly improve performance when predicting top-level segments. Moreover, the fact that the feature OVR has a positive impact on the model for predicting top-level topic boundaries, but does not improve the model for predicting sub-level discourse boundaries reveals that the overlapping speech feature is more predictive of major topic shifts than of subtopic shifts.

6.4.2 Experiment 2: Effects of statistically learned cue phrases

Galley et al. (2003) empirically identified cue phrases that are indicators of major seg-

[7] Because we do not wish to make assumptions about the underlying distribution of error rates, and error rates are not measured on an interval level, we use a non-parametric sign test throughout these experiments to compute statistical significance.

ment boundaries, and then eliminated all cues that had not previously been identified as cue phrases in the literature. Here, we conduct an experiment to explore how different ways of identifying cue phrases can help identify useful new features for the two boundary prediction tasks.

In each fold of the 25-fold leave-one-out cross validation, we use a modified Chi-squared test[8] to calculate statistics for each word (unigram) and word pair (bigram) that occurred in the 24 training meetings. The Chi-squared scores are computed in a way similar to that applied to find decision-characteristic cue phrases in Section 4.2, measuring the association strength between the occurrence of a particular N-gram and that of the discourse segment boundaries.

We then rank unigrams and bigrams according to their Chi-squared scores (i.e. discriminability of discourse segment boundaries), filtering out those with values under 6.64, the threshold for the Chi-squared statistic at the 0.01 significance level. The unigrams and bigrams in this ranked list are the learned cue phrases.[9] Occurrence counts of the cue phrases (EDI-CUE) in an analysis window around each potential topic boundary are then used in the test meeting as a feature.

RANK	1	2	3	4	5	6	7
ALLSEG	LCV	OVR	GAP	**EDICUE**	LCP	**COLCUE**	SP_k
TOPSEG	LCV	SP_k	OVR	LCP	GAP	**EDICUE**	**COLCUE**

Table 6.3: *Ranked list of feature relevance at the TOP and ALL level (in descending order).*

Table 6.3 is a ranked list of feature relevance based on the Chi-squared statistics. It reveals that the statistically selected cue phrases (EDI-CUE) are a more accurate predictor than those that are empirically collected from literature (COL-CUE). The ranked list also confirms the distinctive positive impact of using LF features (LCV and LCP) on predicting discourse segment boundaries.

Table 6.4 shows the performance (P_k) of models that use statistically learned cue phrases in their feature sets compared with models using no cue phrase features and Galley et al. (2003)'s model, which only uses cue phrases that correspond to those

[8] In order to satisfy the mathematical assumptions underlying the test, we remove cases with an expected value that is under a threshold (in this study, we use 1), and apply Yate's correction, $\frac{(|ObservedValue - ExpectedValue| - 0.5)^2}{ExpectedValue}$.

[9] Example cue unigrams selected include 'ok', 'okay', 'but', 'so', 'and', 'yeah', 'um', 'agenda', 'shall', 'items', 'wanted', 'let', 'alright', 'go', 'if', 'that', 'about', 'why', 'ask', 'uh', 'digits', 'you', 'they', 'read', 'um', 'in', 'should', 'the', 'of', 'thing', 'transcription', 'but', 'disk', 'be', 'mikes', 'a', 'know', 'do', 'for', 'mm', and 'good'.

	NOCUE	COL-CUE	1gram	2gram	1+2gram	MC-B	Topline
ALLSEG	0.381	0.369	0.324	0.369	0.350	0.466	n/a
TOPSEG	0.296	0.284	0.290	0.292	0.293	0.484	0.135

Table 6.4: *Performance (P_k) of models trained with cue phrases from the literature (COL-CUE) and cue phrases learned from statistical tests, including cue words (1gram), cue word pairs (2gram), and cue phrases composed of both words and word pairs (1+2gram). NOCUE is the model using no cue phrase features. The Topline is the human annotations on top-level segments.*

identified in the literature (COL-CUE). We see that for predicting all segments, models using the cue word features (1gram) and the combination of cue words and bigrams (1+2gram) yield a 15% and 8.24% improvement over models using no cue features (NOCUE) ($p < 0.01$) respectively, while models using only cue phrases found in the literature (COL-CUE) improve performance by just 3.18%. In contrast, for predicting top-level topics, the model using cue phrases from the literature (COL-CUE) achieves a 4.2% improvement, and this is the only model that produces statistically significantly better results than the model using no cue phrases (NOCUE). The superior performance of models using statistically learned cue phrases as features for predicting finer-grained segment boundaries suggests there may exist a different set of cue phrases that serve as segmentation cues for more subtle topic shifts in discourse.

6.4.3 Experiment 3: Online segmentation from ASR transcripts

Manual transcripts are costly and time-consuming to produce, and thus it is crucial for developing on-line meeting applications that the need for manual transcription is eliminated. Any fully automatic discourse segmentation system that uses lexical features will need to include an ASR system in the initial stage. It is therefore necessary to understand the performance degradation caused by word errors in the transcripts. The experiments in this section examine the degree of degradation in the tasks of predicting segment boundaries at the two levels of granularity. All of the results are reported on the test set.

We repeat the procedure used in Experiment 1 to evaluate segmentation error rates using the LCSeg and combined models on the ASR transcriptions.[10]

[10]We do not report W_d scores for the combined model (CM) on ASR outputs because this model predicted 0 segment boundaries when operating on ASR output. In our experience, CM routinely un-

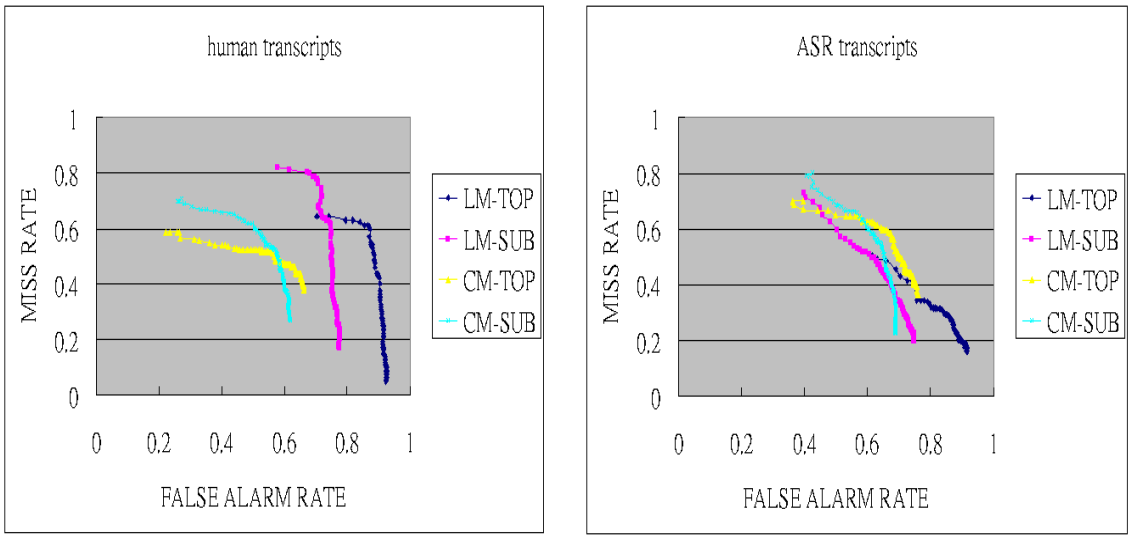

Figure 6.3: *Effects of transcript versions on LCSeg and the combined model when used to predict top-level and all (including subdiscourse) segment boundaries.*

Note that the performance obtained on human transcripts and that on ASR outputs are not directly comparable. This is due to the fact that no word alignment procedure has been applied to ensure that the number of words in the two versions of the transcripts are the same, and thus rendered the word-based segmentation error rate metrics, P_k and W_d not directly comparable. For details of the ASR system in use, please refer to Section 3.1.2.

Features extracted from ASR transcripts are distinct from those extracted from human transcripts in at least three ways: (1) incorrectly recognized words incur erroneous lexical cohesion features (aLF), (2) incorrectly recognized words incur erroneous cue phrase features (aCUE), and (3) the ASR system recognizes less overlapping speech (aOVR).

The rightmost column in Table 6.1 demonstrates the performance of LCSeg and the combined model (CM) when used to segment ASR transcripts. The preferred approaches for segmentation at the two different levels of granularity remains consistent as those for segmentation on human transcripts. For the task of predicting top-level segment boundaries, CM outperforms LCSeg, as LCSeg tends to over-predict at the top-level. For the task of predicting sub-level topic shifts, LCSeg alone is considerably better than CM.

derestimated the number of segment boundaries, and due to the nature of the W_d metric, it should not be used when there are 0 hypothesized topic boundaries.

The DET curves in Figure 6.3 show the effect of the transcript version on detection errors; operating on ASR outputs yields less variation. This shows that the speaker interaction features introduced in the combined model do not make much difference when predicting from ASR outputs. To understand the full potential of the automatic discourse segmentation component, we must further analyze the effect across feature combinations on ASR transcripts.

In contrast to the finding that integrating speaker interaction features with lexical cohesion features is useful for prediction on human transcripts, Table 6.5 shows that, when operating on ASR output, neither adding interaction nor cue phrase features improves the performance of the model using only lexical cohesion features. In fact, the model using all features (ALL) is significantly worse than that using only lexical cohesion features (aLF). This suggests that we must explore new features that can lessen the perplexity introduced by ASR outputs in order to train a better model.

ASR	Error Rate(P_k)	
Feature set	ALLSEG	TOPSEG
MC-B	0.434	0.452
aLF(LCV+LCP)	0.368	0.253
IF(ACT+aOVR+GAP)	0.368	0.253
IF+aCUE	0.368	0.253
aLF+GAP	0.367	0.246
aLF+IF	0.368	0.282
aLF+aCUE	0.374	0.253
ALL(aLF+IF+aCUE)	0.382	0.284

Table 6.5: *Effects of feature combinations for predicting boundaries from ASR output.*

6.5 From ICSI to AMI Meeting Segmentation

Comparison of the ICSI meetings and the AMI meetings (c.f. Section 3.1.1) demonstrates several differences between them. First, the AMI meetings are scenario-driven meetings in which meeting participants were assigned tasks to accomplish in the meeting, while the ICSI meetings are less structured progress report meetings, usually without a clear goal.

Second, the AMI meeting corpus includes video recordings and multi-layer annotations, e.g., dialogue act reflexivity and number of addresses, which do not exist for the ICSI meetings. Consequently, there are more features and more multimodal information integration techniques that can be leveraged.

Third, the progression of the AMI scenario meetings follows a certain structure predetermined in the meeting agenda. Therefore, knowledge about the agenda items commonly seen in these meetings – in other words, "topics" – can lend support to the development of the segmentation models.

Therefore, in the following experiments, in addition to the lexical and speaker interaction features used in the task of ICSI meeting segmentation, we also explore the use of features from the other available knowledge sources, such as prosody, motion, and dialogue context, for the task of recovering discourse structure from multiparty conversational speech. These experiments are designed to answer the following questions: (1) Can the machine learning approach further integrate these potentially characteristic multimodal features for automatic discourse segmentation? (2) What are the most discriminative knowledge sources for detecting segment boundaries in the AMI scenario meetings? (3) Does the use of ASR transcription significantly degrade the performance of a multimodal segmentation model?

In implementing the machine learning approach, we again use the Maximum Entropy (MaxEnt) classifier[11] in the following experiments on the task of AMI meeting segmentation. Because the feature set for these experiments is large, the MaxEnt classifier is used to train models to use only features whose weights can be reliably estimated. Also, as opposed to many other classifiers, MaxEnt does not impose the independence assumption for the probability distribution underlying the training samples. Moreover, previous work has shown MaxEnt models to be effective on the tasks of sentence and discourse segmentation (Liu et al., 2004; Christensen et al., 2005; Hsueh and Moore, 2006). MaxEnt models are also useful in the decision detection task we reported in Chapter 5.

Because continuous features have to be discretized for MaxEnt, we apply a histogram binning approach, which divides the value range into N intervals that contain an equal number of counts as specified in the histogram, to discretize the data. In our experiments, the parameters of the MaxEnt classifier are optimized using Limited-Memory Variable Metrics.

[11] c.f. Section 5.3.1.

6.5.1 Using multimodal and multiparty interaction features

As reported in Chapter 3, there is a wide range of features that are potentially characteristic of segment boundaries. For example, previous research has shown that interlocutors do speak and behave differently when trying to end a discussion and initiate a new one, pause longer than usual when making sure that everyone is ready to move on to a new discussion and use certain conventional expressions (e.g., *well*, *okay*, *let's*) when attempting to get everyone's attention about an upcoming new discussion. We expect to find some of these features useful for automatic recognition of segment boundaries. The features we explore can be divided into the following five classes:

Conversational Features (CONV): As in Experiment 1-3, we extract a set of speaker interaction features, including the amount of overlapping speech, the amount of silence between speaker segments, and speaker activity change. These speaker interaction features, together with the number of cue words and the predictions of LCSeg (i.e., the lexical cohesion statistics, the estimated posterior probability, the predicted class), are grouped into the category of conversational features in the following experiments.

Lexical Features (LX1): Following the "bag of words" representation of documents used for document classification, we back off from high-level descriptions of documents to low-level order-free representations. We compile the list of words that occur more than once in the spurts that have been marked as a top-level or sub-segment boundary in each fold of the training data. Each spurt is then represented as a vector space of unigrams from this list.

Prosodic Features: Prosodic features are suprasegmental features that can be derived from the intonation, rhythm, and lexical stress in speech. Functionally, prosodic features, i.e., intonation, energy, and fundamental frequency (F0), are used to indicate segmentation and saliency (Shriberg et al., 2000; Grosz and Hirschberg, 1992; Liu et al., 2004). As described in Section 5.3.3, we follow Shriberg and Stolcke (2001)'s direct modeling approach to manifest prosodic features, among other things, as duration, pause, speech rate, pitch contour, and energy level, at different points of a spurt.

Prosodic features in context are also considered. As prior research has shown the benefits of including immediate prosodic contexts, this study includes features that provide information about the preceding and following spurts. Table 5.1 contains a list of prosodic context features used in this study.

Motion Features (MOT): We measure the magnitude of relevant movements in the

meeting room using TNO's motion capture system, which detects movements directly from video recordings in frames of 40 ms [12]. Of special interest are the frontal shots as recorded by the close up cameras, the hand movements as recorded by the overview cameras, and shots of the areas of the room where presentations are made. We then average the magnitude of movements over the frames within a spurt as its feature value.

Contextual Features (CTXT): These include dialogue act type[13] and speaker role (e.g., *project manager*, *marketing expert*). As each spurt may consist of multiple dialogue acts, we represent each spurt as a vector of dialogue act types, wherein a component is 1 or 0 depending on whether the dialogue act type occurs in the spurt.

6.5.2 Using phonetic transcription

As noted in the introduction, the phonetic units are successfully used as proxies of word units in many spoken language understanding applications. In Experiment 6, we will attempt to extend the lexical similarity-based approach to work on phonetic transcripts. Specifically, we modify LCSeg to segment AMI meetings by locating dramatic changes between two adjacent windows of phonetic units.

To convert speech signals into a sequence of phonetic units, we employ a phoneme recognition model (Schwarz et al., 2004) that has been successfully applied to multilingual tasks (e.g., automatic language identification (Matejka et al., 2005)) and other spoken language understanding tasks (e.g., speech recognition and keyword spotting). The phoneme recognizer is trained on ten hours of the SpeechDat-E corpus[14], which records spontaneous telephone conversations of 1,000 Hungarian speakers and their pronunciation lexicon. We use the phonotactic model that is trained on the part of Hungarian speaker data in the corpus, because this model outperforms all the other eleven phonotactic models in the NIST2003 Language Identification (LID) task (Matejka et al., 2005). It yielded the lowest error rate on the LID task, even in the face of a high phoneme recognition error rate 41.96%. The recognizer operates in the following three steps.

[12] TNO is one of the sites that collected the AMI corpus. TNO stands for Knowledge for Business in Dutch. c.f. http://www.tno.nl for more information.

[13] In the annotations, each dialogue act is classified as one of 15 types, including acts about information exchange (e.g., "Inform"), acts about possible actions (e.g., "Suggest"), acts whose primary purpose is to smooth the social functioning (e.g., "Be-positive"), acts that are commenting on previous discussion (e.g., "Elicit-assess", "Elicit-inform", "Elicit-suggest"), and acts that allow complete segmentation (e.g., "Stall").

[14] Eastern European Speech Databases for Creation of Voice Driven Teleservices. `http://www.fee.vutbr.cz/SPEECHDAT-E/`.

- **Feature extraction**: First, speech signals are divided into frames of 25 ms long with 10 ms shift. Next, for each frame the system utilizes a Mel-filter bank to obtain its short-term critical band logarithmic spectral density. Finally, temporal pattern (TRAP) feature vectors, i.e., temporal evolution of critical band spectral densities within a single critical band, are generated.

- **Phoneme classification**: For each critical band a neural network classifier is trained to estimate the posterior probabilities of sub-lexical classes (i.e., phonemes). Then, the outputs of these single band classifiers are merged to another neural network classifier such that a combined estimation of phoneme probabilities can be yielded.

- **Representation preparation**: A Viterbi decoder is used to produce phoneme strings. We then organize the sequence of phoneme strings into spurts, i.e., speaker turns with pause no longer than 0.5 seconds in-between.

6.5.3 Using speaker activity information

Previous work has demonstrated that changes in speaker activity are indicative of multiparty discourse segment boundaries (Renals and Ellis, 2003; Galley et al., 2003; Hsueh and Moore, 2007a). In this work we incorporate the following two types of speaker activity into the recognized phonetic transcripts. Figure 6.6 shows the notations used in the transcripts.

The first type (*SPK*) includes speaker movements which are characterized by speaker noises (e.g., lip movement, cough), intermittent noises (e.g., door open, note taking), fillers (e.g., 'hmm', 'ah') and pauses. The phoneme recognizer we use in this work can provide such information.

SPK	spk	speaker noise (lip movement, cough, etc.)
	int	intermittent noise, not from speaker (door, pen, etc.)
	fil	filled voice caused by speaker, e.g., hhmmmm, ahhhh
	pau	pause
ACT	SP_{id}	indicate the previous phoneme is uttered by whom. e.g., SP_b is uttered by the speaker whose id is b.

Table 6.6: *Notations used in phonetically recognized transcripts.*

The second type (*ACT*) depicts how talkative each speaker is over the sequence of

spurts in the phonetic transcripts. Speaker dominance is characterized as the number of phonemes transpired in each spurt by each speaker; accordingly, we could enhance the phonetic transcription with speaker ID tags, SP_{id}, each of which refers to the speaker of a recognized phoneme. Table 6.7 (b) is the speaker activity-augmented version of the phoneme representation in Table 6.7 (a).

(a)	pau int h m o l k S spk s E m h u E k S m u: l k h E S O k S n E n spk pau int n m spk spk o m O k pau int
(b)	pau int h SP_b m SP_b o SP_b l SP_b k SP_b S SP_b spk s SP_b E SP_b m SP_b h SP_b u SP_b E SP_b k SP_b S SP_b m SP_b u: SP_b l SP_b k SP_b h SP_b E SP_b S SP_b O SP_b k SP_b S SP_b n SP_b E SP_b n SP_b spk pau int n SP_b m SP_b spk spk o SP_b m SP_b O SP_b k SP_b pau int
(c)	and uh and he does a funny thing where he chases his tail as well , which is quite amusing , so

Table 6.7: *Example of speaker activity-augmented phonetic representation and its word-based representation.*

6.6 AMI Meeting Segmentation

6.6.1 Experiment 4: Off-line AMI discourse segmentation from human transcripts

The first question we want to address is whether the different types of multimodal and multiparty interaction features can be integrated, using the conditional MaxEnt model, to automatically detect segment boundaries. In this study, we use a set of 50 meetings, which consists of 17,977 spurts. Among these spurts, only 1.7% and 3.3% are top-level and sub-segment boundaries. For our experiments we use 10-fold cross validation. The baseline is the result obtained by using LCSeg, an unsupervised approach exploiting only lexical cohesion statistics. Conventional measures of error rates in segmentation, P_k and W_d, are again used as the evaluation metrics.

Table 6.8 shows the results obtained by using the same set of conversational (CONV) features used in Galley et al. (2003) and this thesis, and results obtained by using all the available features (ALL). The conversational features include pause, overlap ratio, LCSeg score and reported posterior probaility, and speaker activity change. (c.f. Section 6.5.1 for more details.)

	TOPSEG		ALLSEG	
Error Rate	P_k	W_d	P_k	W_d
BASELINE(LCSeg)	0.40	0.49	0.40	0.47
MAXENT(CONV)	0.34	0.34	0.37	0.37
MAXENT(ALL)	0.30	0.33	0.34	0.36

Table 6.8: *Performance comparison of MaxEnt models trained with only conversational features (CONV) and with all available features (ALL).*

In Row 2 of Table 6.8, we see that using a MaxEnt classifier trained on the conversational features (CONV) alone improves over the LCSeg baseline by 15.3% for top-level segments and 6.8% for all-level segments. Row 3 shows that combining additional knowledge sources, including lexical features (LX1) and the non-verbal features, prosody (PROS), motion (MOT), and context (CTXT), yields a further improvement (of 8.8% for top-level segmentation and 5.4% for sub-level segmentation) over the model trained on speaker interaction features. (c.f. Section 6.5.1 for a complete list of features included in the set of CONV, PROS, MOT, and CTXT features.)

The second question we address is which knowledge sources (and combinations) are good predictors for segment boundaries. In this round of experiments, we evaluate the performance of different feature combinations. Table 6.9 further illustrates the impact of each feature class on the error rate metrics (P_k/W_d). In addition, as the P_k and W_d scores do not reflect the magnitude of over- or under-prediction, we also report the average number of hypothesized segment boundaries (Hyp). The number of reference segments in the annotations is 8.7 at the top-level and 14.6 at the sub-level.

Rows 2-6 in Table 6.9 show the results of models trained with each individual feature class. We performed a one-way ANOVA to examine the effect of different feature classes. The ANOVA suggests a reliable effect of feature class ($F(5,54) = 36.1$; $p < .001$). We performed post-hoc tests (Tukey HSD) for significance.

Analysis shows that the model that is trained with lexical features alone (LX1) performs significantly worse than the LCSeg baseline ($p < .001$). This is due to the fact that cue words, such as *okay* and *now*, learned from the training data to signal segment boundaries, are often used for non-discourse purposes, such as making a semantic contribution to an utterance.[15] Thus, we hypothesis that these ambiguous cue words

[15]Hirschberg and Litman (Hirschberg and Litman, 1987) have proposed to discriminate the different uses intonationally.

	TOPSEG			ALLSEG		
	Hyp (SDis)	P_k	W_d	Hyp (SDis)	P_k	W_d
BASELINE (LCSeg)	17.6	0.40	0.49	17.6 (0.21)	0.40	0.47
LX1	61.2 (1.02)	0.53	0.72	65.1 (3.46)	0.49	0.66
CONV	3.1 (-0.64)	0.34	0.34	2.9 (-0.80)	0.37	0.37
PROS	2.3 (-0.75)	0.35	0.35	2.5 (-0.83)	0.37	0.37
MOT	96.2 (10.06)	0.36	0.40	96.2 (5.59)	0.38	0.41
CTXT	2.6 (-0.70)	0.34	0.34	2.2 (-0.85)	0.37	0.37
ALL	7.7 (-0.11)	0.29	0.33	7.6 (-0.48)	0.35	0.38

Table 6.9: *Effects of individual feature classes on AMI discourse segmentation.*

have led the LX1 model to over-predict. Row 7 further shows that when all available features (including LX1) are used, the combined model (ALL) yields performance that is significantly better than that obtained with individual feature classes ($F(5,54) = 32.2$; $p < .001$).

	TOPSEG			ALLSEG		
	Hyp (SDis)	P_k	W_d	Hyp (SDis)	P_k	W_d
ALL	7.7 (-0.11)	0.29	0.33	7.6 (-0.48)	0.35	0.38
ALL-LX1	3.9 (-0.55)	0.35	0.35	3.5 (-0.76)	0.37	0.38
ALL-CONV	6.6 (-0.24)	0.30	0.34	6.8 (-0.53)	0.35	0.37
ALL-PROS	5.6 (-0.36)	0.29	0.31	7.4 (-0.49)	0.33	0.35
ALL-MOTION	7.5 (-0.14)	0.30	0.35	7.3 (-0.50)	0.35	0.37
ALL-CTXT	7.2 (-0.17)	0.29	0.33	6.7 (-0.54)	0.36	0.38

Table 6.10: *Effects of taking out each individual feature class from the ALL model.*

Table 6.10 illustrates the error rate change (i.e., increased or decreased P_k and W_d score)[16] that is incurred by leaving out one feature class from the ALL model. Results show that CONV, PROS, MOTION and CTXT can be taken out from the ALL model individually without increasing the error rate significantly.[17] Moreover, the combined models always perform better than the LX1 model ($p < .01$), cf. Table 6.9. This sug-

[16]Note that the increase in error rate indicates performance degradation, and vice versa.

[17]Sign tests were used to test for significant differences between means in each fold of cross validation.

gests that the non-lexical feature classes are complementary to LX1, and thus it is essential to incorporate some, but not necessarily all, of the non-lexical classes into the model.

	TOP			ALL		
	Hyp (SDis)	P_k	W_d	Hyp (SDis)	P_k	W_d
LX1	61.2 (6.03)	0.53	0.72	65.1 (3.46)	0.49	0.66
MOT	96.2 (10.06)	0.36	0.40	96.2 (5.59)	0.38	0.41
LX1+CONV	5.3 (-0.39)	0.27	0.30	6.9 (-0.53)	0.32	0.35
LX1+PROS	6.2 (-0.29)	0.30	0.33	7.3 (-0.50)	0.36	0.38
LX1+MOT	20.2 (1.32)	0.39	0.49	24.8 (0.70)	0.39	0.47
LX1+CTXT	6.3 (-0.28)	0.28	0.31	7.2 (-0.51)	0.33	0.35
MOT+PROS	62.0 (6.13)	0.34	0.34	62.1 (3.25)	0.37	0.37
MOT+CTXT	2.7 (-0.69)	0.33	0.33	2.3 (-0.84)	0.37	0.37

Table 6.11: *Effects of combining complementary features on AMI discourse segmentation.*

Table 6.11 further illustrates the performance of different feature combinations on detecting segment boundaries. By subtracting the P_k or W_d score in Row 1, the LX1 model, from that in Rows 3-6, we can tell how essential each of the non-lexical classes is to be combined with LX1 into one model. Results show that CONV is the most essential, followed by CTXT, PROS and MOT. The advantage of incorporating the non-lexical feature classes is also shown in the noticeably reduced number of over-predictions as compared to that of the LX1 model.

The column Hyp reported in this table can be used to determine which algorithms result in better approximations in terms of the number of hypothesized segments. Combining any of the non-lexical feature classes reduces the number of over-predictions by LX1 noticeably. Further comparison of performance improvement across the top-level and the all-level segmentation models suggests that little difference exists between these results. However, none of the feature combinations yields a model that is good at estimating the number of all-level segment boundaries.

6.6.2 Experiment 5: Segmentation from ASR transcripts

A major challenge facing the development of an online segmenter is whether the segmentation task can be performed in a fully automatic fashion. Thus we again consider

the performance degradation caused by the ASR recognition errors. Furthermore, as Section 6.6.1 shows that lexical features must be combined with other features to be useful for segmentation, it is also essential to study whether the recognition errors can be compensated for by including other features in the models.

As in Experiment 3, the ASR transcripts used in this experiment are obtained using standard technology including HMM based acoustic modelling and N-gram based language models (Hain et al., 2005). The average word error rate (WER) is 39.1%.

Because the P_k and W_d scores are computed as the number of prediction errors normalized by the number of words in the reference transcript, prediction algorithms that have made the same number of errors can yield dfifferent P_k and W_d scores. In other words, the comparison across the scores obtained in different rounds is only fair when the reference transcripts consist of the same number of words. To compare the P_k and W_d metrics obtained on the ASR outputs directly with those obtained on the human transcripts, a word alignment algorithm has been applied to ensure each word in the ASR transcripts has one corresponding word in the manual transcripts, so that the number of words in the ASR and the manual transcripts will remain the same.

The word alignment algorithm used by the AMI grop works as follows: First, the timings were generated using acoustic models of an automatic speech recog- nition system reported in Section 3.1.2. The system was specifically developed for the transcription of the AMI meetings using all input channels, with aids from the Hidden Markov Model Toolkit (HTK). The time level information it- self was obtained in a multi-step process.

Then, a viterbi alignment procedure is applied to encode the timing information into the acoustic recordings from the independent headset microphones. Utterance time boundaries are used from the previous segmentation. Two passes of alignment are necessary to ensure a fixed silence collar for each utterance. The output of the above process is an exact time and duration for each pronounceable word in the corpus according to close talking microphones.

Finally, word level times should be broadly correct, however problems arise in the vicinity of overlapped speech (i.e. multi- ple speakers talking at the same time) and non-speech sounds (like door-closing etc). For more details, please refer to Carletta et al. (2006).

In this study, we again use the set of 50 meetings and 10-fold cross validation. we compare the performance of the reference models, which are trained on human transcripts and tested on human transcripts, with that of the ASR models, which are

trained on ASR transcripts and tested on ASR transcripts. Table 6.12 shows that despite the word recognition errors, none of the LCSeg, the MaxEnt models trained with speaker interaction features, or the MaxEnt models trained with all available features perform significantly worse on ASR transcripts than on reference transcripts. One possible explanation for this, which we have observed in our corpus, is the consistency of recognition errors. The ASR system is likely to misrecognise different occurrences of words in the same way, and thus the lexical cohesion statistic, which captures the similarity of word repetition between two adjacency windows, is likely to remain unchanged. In addition, when the models are trained with other features (e.g., pause) that are not affected by recognition errors, the negative impacts of recognition errors are reduced to an insignificant level.

	TOPSEG		ALLSEG	
Error Rate	P_k	W_d	P_k	W_d
LCSeg(REF)	0.45	0.57	0.42	0.47
LCSeg(ASR)	0.45	0.58	0.40	0.47
MAXENT-CONV(REF)	0.34	0.34	0.37	0.37
MAXENT-CONV(ASR)	0.34	0.33	0.38	0.38
MAXENT-ALL(REF)	0.30	0.33	0.34	0.36
MAXENT-ALL(ASR)	0.31	0.34	0.34	0.37

Table 6.12: *Effects of word recognition errors on AMI discourse segmentation.*

6.6.3 Experiment 6: Segmenting directly over audio signals

In previous experiments, we explored machine learning and time series analysis approaches to the task of meeting discourse segmentation. We have also shown that the multimodal features improve the performance of the machine learning approach. However, we have not studied extensively how to accommodate the multimodal characteristics of meeting discourse in the time series analysis approach like LCSeg. Therefore, in this experiment, we address the challenge of segmenting meeting recordings directly from the audio source, leveraging the phonetic transcripts that can be produced in near real time.

Three versions of transcripts are examined in this study: the lexical transcript (LC), phonetic transcript (PH), and speaker activity-enhanced phonetic transcript (PH+ACT). Table 6.13 demonstrates the effect of transcript version on segmentation performance.

	K				unK					
	TOP		ALLSEG		TOP			ALLSEG		
Transcript Type	P_k	W_d	P_k	W_d	P_k	W_d	SDis	P_k	W_d	SDis
LC	0.36	0.38	0.36	0.40	0.44	0.55	1.11	0.40	0.49	0.42
PH	0.42	0.43	0.43	0.45	0.40	0.41	0.14	0.41	0.42	-0.23
PH+ACT	0.36	0.39	0.40	0.44	0.35	0.36	-0.38	0.39	0.40	-0.58

Table 6.13: *Effects of using speaker activity-enhanced phonetic transcripts on unsupervised segmentation.* P_k *and* W_d *measure the segmentation error rates.* SDis *measures the structural similarity of a hypothesized segmentation to the reference segmentation. The closer to zero the more similar to the reference segmentation.*

Row 1 shows the performance of the LC version, which locates changes over lexical patterns. Rows 2 and 3 show the performance of the PH version, which locates changes over sublexical patterns[18], and that of the PH+ACT version, which locates changes over the sublexical patterns and speaker activities.

In search for segmentation systems that can work well in the online scenario, in this experiment two conditions are attempted: In the first condition we force the segmenter to predict the same number of segments as the ground truth segments (hereafter, the *K* condition)[19]. In the second condition (hereafter, *unK*), we use an empirically determined threshold to select the most probable segment boundaries, without constraining how many to be selected. Our system follows previous work to select only the potential boundary sites whose posterior probabilities are above the average number of segments in a meeting minus half the standard deviation. The first four columns illustrate the *K* condition. Results show that, when the number of segments is given, the LC model does perform better than the PH model. However, when patterns in speaker dominance (ACT) are jointly considered along with phonetic chains, the new PH+ACT model yields competitive performance to the LC model in the task of recovering top-level segments (TOP) in a dialogue structure.

The right six columns illustrate the *unK* condition wherein the number of ground truth segments is unknown. Comparing the results across the two conditions, *K* and *unK*, clearly shows a negative effect of the added structural uncertainty on the LC

[18]The phonetic transcripts include both phonemes and information about speaker movements as explained in Table 6.6.

[19]We experiment with this condition because we want to compare with many of the previous work that use this setting.

model, increasing the error rate[20] by 22% and 11% on recovering segments at the top level and at all levels respectively. In contrast, the added uncertainty does not significantly affect the performance of the PH model. For the task of recovering the top-level segments, the PH model outperforms the LC model by 10%; Adding the model of speaker dominance (PH+ACT) further reduces the error rate by 14%.

These results suggest that speaker activity-related models have greater potential to be used in online applications to recover sub-discourse dialogue segments. Also, as functional segments covers nearly half of the top-level segments (see Section 3.1.1), we expect the accuracy of predicting functional segments to be important to the success of the models for top-level segmentation. Therefore, we perform subsequent experiments to examine the effects of speaker activity-based information on the accuracy of functional segment predictions.

Line 1-3 in Table 6.14 show the results of operating the system on the three versions of transcripts: lexical transcripts (LC), phonetic transcripts (PH), and speaker activity-enhanced phonetic transcripts (PH+ACT). Line 4-5 show the results of locating changes in speaker movements and in speaker dominance respectively. Line 6 shows the result of locating changes in both of these two types of speaker activity information. Results suggest that, when the number of segments is given, all the systems that locate changes in speaker dominance patterns (i.e. ACT, PH+ACT, SPK+ACT) yield better recall than LC. Under the more realistic condition where the number of segments is unknown, these systems yield slightly worse recall than LC; but considering that the models yield good structural similarity (as shown by the low structural distance), they are still a more accurate predictor for functional segments.

The columns denoted as SDis in Table 6.13 and Table 6.14 demonstrate the level of structural similarity between the predictions that are obtained on the ground truth and the different versions of transcripts. The close-to-zero figures of the predictions produced by the ACT-related models (such as PH+ACT, ACT, and SPK+ACT) indicate that speaker activity information is a better predictor of the off-topic functional segments.

[20]Since the scores of P_k and W_d are both aggregated measures of segmentation error rate, we report the change in only P_k.

	K-TOPSEG	K-ALLSEG	unK	
Accuracy/SDis	Recall	Recall	Recall	SDis
LC	0.75	0.78	0.83	6.14
PH	0.65	0.70	0.69	1.91
PH+ACT	0.86	0.88	0.77	0.09
SPK	0.62	0.65	0.61	-1.00
ACT	0.84	0.84	0.77	-.0.05
SPK+ACT	0.82	0.88	0.80	0.39

Table 6.14: *Effects of speaker-activity models on the accuracy of off-topic, functional segment prediction. Under the K-TOPSEG condition, the total number of the ground truth segments at the top level (K_{top}) is given as a constraint to the segmenter for selecting a list of top K predictions from the hypotheses. Under the K-ALLSEG condition, the total number of segments in the two-layer ground-truth segment structure (K_{all}) is given. Under the unK condition, the total number of segments is unspecified.*

6.7 Discussion

Overall, the goal of this research is to identify effective ways to segment meetings. In this chapter, we made considerable progress towards achieving the goal by addressing these questions.

1. How to adapt the methods previously developed in text and broadcast news segmentation to segment meetings, integrating the various potentially predictive knowledge sources?

2. What are the effective knowledge sources serving to find discourse segments of different natures?

3. How to adapt the offline segmenters to be operated online or right after the end of a meeting?

6.7.1 How to adapt the methods previously developed in text and broadcast news segmentation to segment meetings?

As observed in the meeting corpora, the lack of many macro-level discourse units that are useful for segmenting text or broadcast news, e.g., story and paragraph breaks, has posed new challenges to the task of segmenting meeting dialogues. Furthermore, many

discourse information are implicit in the multimodal and multiparty cues conveyed by the meeting participants. This has rendered the unsupervised time series analysis approaches like LCSeg, which focus on measuring lexical similarity between adjacent windows, less effective on the task of meeting discourse segmentation.

Consequently, exploring novel ways that can capture patterns in the multimodal and multiparty cues is necessary. In keeping with this attempt, we have explored the use of phoneme information, which can be obtained directly from audio inputs in near real time. The figures in Table 6.13 confirmed that the phoneme-based segmentation (PhSeg) approach can secure the same level of segmentation accuracy, while improving greatly on efficiency – from 30 times real time (estimated) to near real time. Enhanced with information about speaker activity, PhSeg can further improve the performance of the unsupervised approach on the task of recovering the top-level discourse structure by 12.5%.

Also observed in the meeting data, the naturally imbalanced class distribution of segment boundaries and non-boundaries has posed challenges to the development of a well-performing supervised machine learning model. Compared to the task of segmenting expository texts, which has been reported in Hearst (1997) with a **39.1%** chance of each paragraph end being a target topic boundary, the chance of each speaker turn being a top-level or sub-level discourse segment boundary in the ICSI meetings is just **2.2%** and **0.69%**. Therefore, it is essential to study methods to cancel off the negative impact brought by the imbalanced distribution.

One strategy to cope with the imbalanced class distribution is to apply **sampling techniques** that compensate the imbalance class distribution in the training set. In a pilot study, we found that sampling techniques previously reported in Liu et al. (2004) as useful for dealing with an imbalanced class distribution in the task of disfluency detection and sentence segmentation do not work for the ICSI meeting data set. The implicit assumption of some classifiers (such as pruned decision trees) that the class distribution of the test set matches that of the training set, and that the costs of false alarms and missed detections are equivalent, may account for the failure of these sampling techniques to yield improvements in performance, when measured using P_k and W_d.

Another coping strategy that does not change the natural class distribution is to **increase the size of the training set**. We conducted an experiment in which the training set size were incrementally increased by randomly choosing ten meetings each time until all meetings were selected. We executed the process three times and averaged

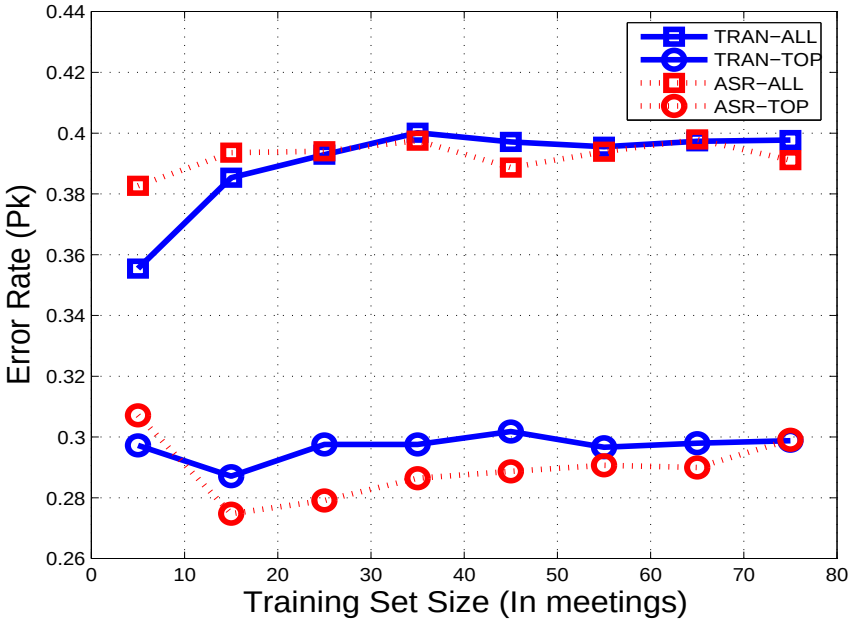

Figure 6.4: *Learning curve of the combined model over the increase of the training set size.*

the P_k scores to obtain the results shown in Figure 6.4. This procesure was applied to examine the effect of training set size in the following dataset: (1) the segmentation at the top-level (TRAN-TOP) and at the sbtopic level (TRAN-ALL) on manual transcripts, and (2) the segmentation at the top-level (ASR-TOP) and at the sbtopic level (ASR-ALL) on ASR transcripts.

Results show that increasing training set size decreases the top-level segmentation error rate to a small extent. (Whether the difference is statistically significant is untested.) But the performance will not receive any more notable improvement after the training set size has reached fifteen. The figure indicates that training a model to predict top-level boundaries requires no more than fifteen meetings in the training set to reach a reasonable level of performance. One conjecture is that the extra meetings may have added to the perplexity of model computation in the training phase.

In contrast, the error rates of sub-level discourse segmentation increase with the increasing training set size till the set reaches thirty-five meetings. This is inconsistent with our hypothesis. It seems that the heterogeneous types of segments – e.g., the activity-based and the topic-based types – inherent in the sub-level discourse structure have further degraded the perplexity when more meetings are introduced in the training set.

In sum, when adapting the text or broadcast news segmentation to segment meeting

speech, it is important to incorporate more knowledge sources such as phoneme and speaker activity information that can capture the implicit multimodal and multiparty cues in conversations. Also, when the supervised learning approach is adopted, it would be beneficial to find out the lower bound of the amount of data needed to balance out the negative effect of an imbalanced class distribution.

6.7.2 What are the effective knowledge sources serving to find discourse segments of different types?

In the previous experiments (Experiment 1 and 4), we have explored the use of many different knowledge sources, ranging widely from those used in previous work, e.g., conversational cues, to features that have not been attempted, e.g., statistically selected cue words, prosody, motion, dialogue context.

There exists at least one difference in the feature effects for AMI meeting segmentation and those for ICSI meeting segmentation. Contradictory to the finding on the AMI meetings[21], the unsupervised LCSeg algorithm that uses solely lexical cohesion information is better for predicting sub-level discourse boundaries. This tells us the sub-level discourse segments in the ICSI meetings involve more topic shifts than the rest of discourse segment types.

A closer examination of the effect of the different knowledge sources sources on segmentation accuracy has suggested the following.

1. To achieve the best performance, the lexical features need to be combined with some, if not all, of the other multimodal or multiparty interaction feature classes. The advantage of incorporating these feature classes comes from their capabilities of reducing number of over-predictions resulted from the inclusion of lexical features.

2. The positive effect of lexical features is more pronounced when the task is to find discourse segments that involve topic shifts, e.g., those at the sub-level of the ICSI meeting discourse structures.

3. The multiparty interaction features, i.e., the speaker interaction and dialogue context features, can significantly improve the performance of the lexical models when being combined. The positive effect of the interaction features are

[21] For the task of AMI discourse segmentation, the combined model works better for predicting boundaries at both the top- and sub-level of discourse.

observed especially clearly at the top-level of the ICSI meeting structures and the both layers in the AMI structures.[22]

4. Among the multimodal features, only the prosodic features are beneficial, and its advantage is limited to finding the top-level discourse segments that are more activity-based.

Having established the benefits of including additional knowledge sources for recognising segment boundaries, the next question to be addressed is **what features in these sources are most useful for recognition of segment boundaries**. To provide a qualitative account of the segmentation cues, we performed an analysis to determine whether each proposed feature discriminates the class of segment boundaries. Previously, in (Hsueh and Moore, 2006) we have shown how the statistical measures (e.g., Log Likelihood ratio) are useful for determining the statistical association strength (relevance) of the occurrence of an n-gram feature to target class. Here we extended that study to calculate the Log Likelihood relevance of all of the features used in the experiments, and use the statistics to rank the features.

Our analysis has shown that people do speak and behave differently near segment boundaries. Some of the identified segmentation cues match previous findings. For example, a segment is likely to start with higher pitched sounds (Brown et al., 1980; Ayers, 1994) and a lower rate of speech (Lehiste, 1980). Also, interlocutors pause longer than usual to make sure that everyone is ready to move on to a new discussion (Brown et al., 1980; Passonneau and Litman, 1993) and use some conventional expressions (e.g., *now*, *okay*, *let's*, *um*, *so*).

Our analysis has also identified segmentation cues that have not been mentioned in previous research. For example, interlocutors do not move around a lot when a new discussion is brought up; interlocutors mention agenda items (e.g., *presentation*, *meeting*) or content words more often when initiating a new discussion. Also, through the analysis of current dialogue act types and their immediate contexts, we have also realized that at segment boundaries interlocutors do the following more often than usual: start speaking before they are ready ("Stall"), give information ("Inform"), elicit an assessment of what has been said so far ("Elicit-assess"), or act to smooth social functioning and make the group happier ("Be-positive").

[22]The effect of the dialogue context features is unconfirmed in the ICSI meeting, as we did not use the dialogue context features in the ICSI meeting segmentation experiments.

6.7.3 How to adapt the offline segmenters to be operated online or right after the end of a meeting?

To address this question, we have attempted to develop online segmenters with two techniques: (1) Replacing manual (word-based) transcripts with ASR (word-based) transcripts, and (2) replacing manual (word-based) transcripts with automatically generated sub-lexical (phoneme-based) transcripts.

To evaluate the performance of the first technique, we repeated the experiments that were operated on manual transcripts, with the manual transcripts replaced by ASR outputs. The figures in Table 6.5 and the DET curves in Figure 6.3 have shown that when operating on ASR output, neither adding interaction nor cue phrase features is useful for compensating the performance degradation caused by the less accurate, ASR word-based lexical cohesion features. In other words, it has suggested the need for exploring new features that can lessen the perplexity introduced by the ASR outputs in order to train a better model.

In Experiment 5, we explored the use of additional knowledge sources in a different task, AMI meeting discourse segmentation. The figures in Table 6.12 have demonstrated that the use of ASR transcripts does not degrade the performance of AMI meeting segmentation significantly. Also, the effective knowledge sources serving for the segmentation tasks at different levels of granularity are the same across all transcript versions available in the AMI corpus, regardless it is from ASR outputs or manual annotations. This has indicated that the integration of multimodal and multiparty interaction features can compensate for recognition errors.

To evaluate the performance of the second technique, we explored a novel way to capture lexical patterns, that is, to convert the audio inputs into a sequence of phonetic strings and to derive sub-lexical patterns therein. In addition, we also explored two ways to model non-lexical patterns that pertain to speaker activities: speaker movement (i.e., speaker and intermittent noise, filter, pause) and speaker dominance. A series of models have been developed in Experiment 6 to examine the effectiveness of these different patterns, which can be derived from the audio recordings, at least, in near real time, on the task of recovering a two-layer structure of the AMI meeting discourse. The results have shown that (1) the automatically generated sub-lexical (phonetic) transcripts can replace the manual word transcripts without degrading the performance of segmenters, (2) the speaker activity information are an even better predictor for off-topic functional segments, which cover nearly half of the top-level

segments of the AMI meetings, and (3) it is possible to incorporate the speaker activity information into the sublexical transcripts to further improve the performance of the online unsupervised segmenters.

6.8 Summary and Limitation

Discovering major discourse segments and finding nested sub-segments are essential for the success of spoken document browsing and retrieval. Meeting records contain rich information, in both content and conversation behavioral form, that enable automatic topic segmentation at different levels of granularity. This study explores the use of features from multiple knowledge sources (i.e., words, prosody, motion, interaction cues, speaker intention and role) for developing an automatic segmentation component in spontaneous, multiparty conversational speech. In particular, we have addressed the following questions: (1) Can the methods previously developed for text and broadcast news segmentation be adapted to segment multiparty dialogue recordings, in particular those of small group meetings? (2) What are the most discriminative knowledge sources for detecting segment boundaries? (3) How to adapt the offline segmenters to be operated fully automatically?

In this chapter, we have first focused on finding ways to answer the first question. Among all, we have explored the extensions of an unsupervised time series analysis and a machine learning approach. we have also examined the effectiveness of incorporating a variety of knowledge sources that are expected to be characteristic of discourse segment boundaries.

A series of experiments have been conducted on both the recordings of the natural ICSI meeting and those of the AMI scenario meetings. Results show that for extending the unsupervised time series analysis approach, it is important to incorporate sublexical (phoneme) and speaker activity information that can capture the multimodal and multiparty cues implicit in the conversations. These information can be extracted fully automatically from the audio inputs. Hence, incorporating them has prevented the use of word-based transcripts, which are costly to obtain either manually or automatically, and in turn enabled the development of online segmenters.

Results further show that for extending the supervised learning approach, lexical features would need to be combined with at least some of the other multimodal or multiparty interaction features to achieve the best performance. Also, it would be beneficial to find out the lower bound of the amount of data needed for canceling

out the negative effect of the imbalanced class distribution represented in the data. Our results have improved on previous work, which uses only conversational features, by 8.8% for top-level segmentation and 5.4% for sub-level segmentation in the AMI meetings.

The advantage of incorporating the multimodal and multiparty interaction feature classes comes from their capabilities of reducing number of over-predictions that are resulted from the inclusion of lexical features. The new features can be extracted from knowledge sources that are either evidenced as useful in previous work, e.g., lexical and conversational features, and those that have not been attempted before, e.g., prosody, motion, dialogue context. There are several reasons for the benefits offered by these non-lexical knowledge sources. First, the presence of the non-verbal features in the model can balance off the over-fitting tendency of models trained with lexical cues. Second, because there is an interaction effect between these non-verbal features, by combining these features we can further improve the performance of the segmentation models.

The examination of feature effects has further demonstrated the effective knowledge sources serving for finding different types of discourse may be different. Take ICSI meeting segmentation for example, the knowledge sources that are effective for identifying speaker interaction changes at the top-level and topic shifts at the sub-level segments are distinctly different in many ways: (1) for predicting sub-level segments, using solely the lexical cohesion information achieves results that are competitive with the machine learning approach that combines lexical and conversational features; (2) for predicting top-level segments, the machine learning approach performs the best; and (3) many conversational cues, such as overlapping speech and cue phrases discussed in the literature, are better indicators for top-level discourse segment boundaries than for the topic shifts at the sub-level; (4) but new features such as cue phrases can be learned statistically to improve the performance in the subtopic prediction task.

Take AMI meeting segmentation as another example. The effective knowledge sources for predicting both layers of the AMI discourse structure are mostly similar: Combining all available features works the best for both; the multiparty interaction features, i.e., the speaker interaction and dialogue context features, can significantly improve the performance of the lexical models when being combined. But there still exists some differences in the effect of the multimodal features: For one, the prosodic features are only beneficial for finding the top-level discourse segments but not the subtle ones.

One drawback of many of these approaches lies on the fact that they require word-based transcripts for the extraction of lexical features. Because we do not wish to make the assumption that high quality manual transcripts of meeting records will be commonly available, we have explored the performance degradation caused by operating directly on automatic speech recognition (ASR) output. Results have encouragingly shown that it is possible to segment meeting speech directly on the ASR outputs, by incorporating other knowledge sources.

However, the use of ASR outputs is still not enough for the development of a fully automatic segmentation component in multimedia archives. We need approaches that can segment a meeting when still in progress, since we expect this to be important to the development of downstream online applications that require immediate content-based access. But generating ASR outputs is an efficient task, which estimatedly costs 30 times real time to complete. Therefore, when the ASR outputs are not readily available, we need to have a substitute plan.

To provide such plan, in this work we have explored a novel way to capture lexical patterns, that is, to convert the audio inputs into a sequence of phonetic strings and to derive sub-lexical patterns therein. In addition, we have explored two more ways to model the non-lexical patterns that pertain to speaker activities: speaker movement (i.e., speaker and intermittent noise, filter, pause) and speaker dominance. Experiments have shown that, when the number of discourse segments are known, our audio-based system which consider all of the phonetic and speaker activity-related patterns can yield results comparable to those obtained by operating the system on manual transcripts. Moreover, in the real-life scenario wherein one has missed the first part of a meeting and do not know how many topics have been discussed, our audio-based systems can even significantly outperform those word-based systems.

Results are encouraging as it shows that speaker activity-augmented phonetic units can serve as proxies of words in unsupervised segmentation of meeting dialogues. Our audio-based system can segment meeting dialogues in absence of manual and high quality ASR transcripts. It is desirable to the development of segmentation components that have to be operated online, and even in unfamiliar domains and languages. Also, as the automatically derived dialogue structures can make up for the lack of explicit orthographic cues (e.g., story and paragraph breaks), the audio-based system is expected to be beneficial to developing the online version of many downstream spoken language understanding applications, such as anaphora resolution information retrieval (e.g., as inputs for the TREC Spoken Document Retrieval (SDR) task), summarization,

and machine translation.

There are still several possible extensions of the current approaches we have not touched on in this chapter. First, with the segmentation models developed and discriminative knowledge sources identified, a remaining question is whether it is possible to automatically select the discriminative features for recognition. This is particularly important for prosodic features, because the direct modeling approach we adopted has resulted in a large number of features. It is expected that by applying feature selection methods we can further improve the performance of automatic segmentation models. As shown in Chapter 4, in the field of machine learning and pattern analysis, many methods and selection criteria have been proposed. A natural next step will be to examine the effectiveness of these methods for the task of automatic segmentation.

Also, in previous work, the approaches of automatic knowledge source selection have been proven to be useful to the accuracy of predictions. It will be interesting to further explore how to construct well-performing ensembles from libraries of models generated with different knowledge sources.

Finally, the current approach only consider information from within the analysis windows that are immediately preceding and following each potential segment boundary site. As longer-range dependencies have been shown as useful in the decision detection task (c.f. Chapter 5) and in previous segmentation research (Blei et al., 2003), it will be beneficial to explore models that take into account these longer range dependencies.

Chapter 7

Task-Oriented Evaluation of Meeting Decision Detector

7.1 Introduction

Standard meeting browsers, which come with typical information retrieval and playback facilities, help answer less than 20% of user queries (Pallotta et al., 2007a). This has led researchers to augment meeting browsers with additional interfaces. For example, interfaces that display thematic (i.e., topic) and contextual (e.g., speaker role, meeting state) representations have been found to be useful for meeting information retrieval, helping users find answers in 25% less time (Banerjee et al., 2005). The best automatic meeting summaries were shown to be those that encapsulate answers to the most frequently asked questions (Erol et al., 2003). In this thesis, we facilitate current meeting browser development by developing automatic mechanisms that identify decisions, which have been suggested as the most sought-after information category and the most essential outcome of meetings (Romano and Nunamaker, 2001; Post et al., 2004; Banerjee et al., 2005; Rienks et al., 2005; Whittaker et al., 2005; Pallotta et al., 2007a).

Thus far, we have evaluated our meeting decision detection models intrinsically, using standard metrics. However, we still do not understand how well the detected decision points will fare when used in a real-life task. In this chapter, we describe an extrinsic task-based evaluation of these models. Specifically, we investigate into the effectiveness of these models on a meeting documentation task that is common in organization contexts – we evaluate the use of the decision-focused extractive summaries produced by the meeting decision detector (as introduced in Chapter 5) for finding

meeting decisions from the archives and absorbing them. On average, these models select 1% of the meeting recordings in the extracts. Compared to the general-purpose extracts, which represent compressions of 30-40%, these models produce much shorter and focused summaries. Our conjecture is that the users will leverage the much focused extractive summaries presented to them along with other information in the browser to produce an abstractive summary of all the decisions discussed in the meeting recordings more effectively.

7.2 Related Work

7.2.1 Extractive Summarization: General v.s. Query-Driven Approach

To save time and human labor in generating meeting minutes, techniques have been proposed to produce extractive summaries by distinguishing the informative dialogue units from the uninformative ones in meetings. Traditionally, the extractive technique works well in text summarization: Mani and Bloedorn (1998) have found that users absorb information in summaries faster than in full text, despite some loss of accuracy. As each sentence in an extractive summary may trigger a follow-up action, user interface studies have also provided evidence on how the extracts are beneficial to the information gathering process (Bates, 1990), especially for those who are unfamiliar with the target documents.

Text summarization commonly uses lexical information, such as the counts of cue phrases, co-occurrences, and tf*idf scores (or its variants), to rank the extract-worthiness of each unit (Edmundson, 1968; Kupiec et al., 1995; Teufel and Moens, 2002). Some approaches rely on orthographic cues (e.g., the position and title) and semantic information (e.g., the degree of connectedness in semantic graphs, co-references) (Mani and Bloedorn, 1998; Barzilay, 2003). However, not all of this information is encoded in speech. Past research in speech summarization overcomes this problem by accommodating other types of speech-specific information. For example, some approaches combine lexical and prosodic information to perform summarization in a variety of speech genres, such as broadcast news (Koumpis et al., 2001) and voice mail (Maskey and Hirschberg, 2005).

Extractive techniques have also been applied to identify generically informative units that are reflective of overall meeting content (Zechner, 2002; Murray et al., 2005;

Miekes et al., 2007). Murray et al. (2006) leveraged prosodic information to identify the most informative meeting dialogue acts. Murray (2007) also evaluated an interface that displays the automatically generated extractive summaries and showed that this interface is rated well by users who were asked to audit how a group come to a particular decision (henceforth the "decision audit" task).

Despite their effectiveness in the decision audit task, the often-lengthy general-purpose extractive summaries may not satisfy users who have missed the meetings and needed a quick overview of all the information relevant to a particular query. Instead, we expect an interface that displays query-focused summaries, which are filtered by relevance to the query, would be more helpful. For example, for users who want to obtain an overview of all the decisions made in the previous meetings, navigating through a decision-focused summary first would help. This requires a pre-index step to highlight the decision-related parts of meetings.

7.2.2 Extractive summary evaluation

Whether the summary is general or query-driven, past research on automatic extractive summarization relies on intrinsic measures such as precision and recall to refine the algorithms. Sparck-Jones and Gallier (1996); Zhu and Penn (2005) provide an extensive review of metrics used for comparing extractive summaries with their human-produced counterpart. Current research examines the relations between the intrinsic measures and human judgements. For example, Lin and Hovy (2003) and Furui et al. (2005) have shown that human judgement scores consistently correlate with ROUGE, a recall-based statistic of n-gram co-occurrence[1]. However, in a smaller scale study of meeting speech summaries, Murray et al. (2006) did not find a consistent correlation in this context.

Nenkova et al. (2007) proposed the Pyramid method to address the variation in human summaries. In Pyramid, multiple (usually 4 to 5) model human summaries are obtained, and each summary content unit is weighted by the level of agreement among the model summaries.

The correlation between the Pyramid scores and the human judgements of meeting speech summaries has yet to be addressed due to the labor-intensiveness of this approach. In addition, none of the proposed intrinsic evaluation measures have been evaluated against user performance in real-life tasks. It is therefore uncertain to what

[1]The forebear of ROUGE is BLEU, a precision-based statistics of n-gram occurrence which is commonly used in the evaluation of machine translation algorithms.

extent the automatic summaries, refined by intrinsic metrics, are beneficial to actual user performance. Therefore, in this thesis, we perform an extrinsic evaluation scheme to examine the use of different types of meeting summarization techniques for a common real-life task.

7.2.3 Extrinsic evaluation

Previous research has performed extrinsic evaluations to identify systems useful for retrieving information from multimedia documents. The genres of the targeted documents range widely from audio (e.g., broadcast news (Brown et al., 1995), voice mails (Hirschberg et al., 1998)) to video (e.g., TV news Hauptmann and Witbrock (1997), video mails (Jones et al., 1997)). Many of these evaluated systems aid the retrieval process by incorporating a visual representation that displays indexed documents.

In the context of recorded conversations and video conferences, naturalistic studies have been first conducted to examine the use of hand-written notes, which can be seen as manual indices provided by the meeting participants, in generating summaries (Whittaker et al., 1994; Moran et al., 1997; Banerjee and Rudnicky, 2007). Kazman et al. (1996) implemented a meeting browser, Jabber, to evaluate various indexing techniques extrinsically and showed that it helps meeting participants find information more effectively than simply answering queries from memory. Whittaker et al. (2008) further performed extrinsic evaluations to examine how useful a similar browser, JFerret (Wellner et al., 2004), is for meeting non-participants. This study shows that fact-based questions are relatively easier for the non-participants to answer than general gist questions. To assist those who need to absorb the overall discussion in meetings, Murray (2007) incorporated a general-purpose summary display into the meeting browser. With the additional extractive summary information, the non-participant users were found to perform well in the "decision audit" task, i.e., to provide a summary that describes the argumentative process of a particular decision.

Following the encouraging direction, we pursue a new direction, which is inspired by previous research in query-driven summarization and multimedia indexing to create a visual representation of only those units that help fill in a query-based template. Typically, the queries are specified by the users. In this thesis, we automate the query generation process to capture information relevant to one particular type of queries, that is, decisions. In particular, we develop an interface that displays only the decision-focused summaries (which we call "decision-focused summary display"). In the fol-

lowing sections, we provide an account of the performance of the general-purpose summary display and the decision-focused one in a real-life task.

7.3 Methodology

7.3.1 Task overview

As pointed out in many organizational studies, obtaining an overview of the decisions made in previous meetings is critical to the preparation of future meetings (Pallotta et al., 2005; Rienks et al., 2005; Whittaker et al., 2005). We thus chose a "decision debriefing" task in this study to compare the two types of extractive summaries. The goal of this task is to summarize all the decisions made in a series of meetings. The decision minutes (abstractive summaries) generated by the subjects in the decision debriefing task are different from the decision-focused extractive summaries we present to the subjects in the way that these minutes are expected to be more readable and contain enough details for human interpretation.

In this study, our goal is to investigate into the effects of three major factors: General extractive summary vs. query-focused extractive summary, manual summary vs. automatic summary, reference transcripts vs. automatic transcripts. In turn, we arranged four conditions to test the three factors, each with one factor different from its counterpart condition. Each participant is randomly assigned to one of the four conditions: AE-ASR(using extractive summaries displayed on ASR transcripts), AD-ASR(decision-focused summaries on ASR transcripts), AD-REF(decision-focused summaries on manual transcripts), MD-REF(manual decision-focused summaries displayed on manual transcripts). Later in the data analysis stage, the performance of those subjects given the first condition would be compared with those given the second to understand the effect of summary type (query-focused or not). Likewise, the second and the third condition would be compared to understand the effect of the errors introduced by automatically generated focused summaries; the third and the fourth condition for the effect of the errors introduced by ASR transcripts. More details are in Section 7.3.8.

We recruited 35 subjects (20 females and 15 males, ages from 18 to 44) during a two-month period in 2008 to perform this task.[2] These subjects were recruited from

[2]We recruited 40 subjects initially, 10 for each condition. Due to a memory leak problem in the logging program, incomplete data were logged in the sessions of five subjects: three in the AD-REF condition, one in the AD-ASR condition, and one in the BASELINE condition. Therefore we excluded the data of these five subjects.

the undergraduate and graduate program of distinctively diverse fields (e.g., history, medicine, chemistry, geography). They were asked to fill in a pre-questionnaire (c.f. Appendix A) about their prior experience in computer use and meeting attendance. An experimenter then guided the subject through the procedure. Depending on the condition they were assigned to, they were presented instructions with some slight differences in word choices (c.f. Appendix B and Appendix C).

Subjects were asked to go through one AMI meeting series and to debrief the decisions for their upper management. The four meetings in the series (each lasts 30-40 minutes) are displayed in parallel so that the subjects could easily jump to the meeting recording they were interested in. For readers' information, following is a synopsis of the decision summaries. Note that the original summaries are in the form of bullet points (as exemplified in Figure 7.2).

- In the kick-off meeting (*A*), the entire group decided that the prototype design should be simple, keeping the everyday functions on one interface and more complicated functions on another;
- In the conceptual design (*B*) and detailed design (*C*) meetings, the group decided on the specific target group, the essential functions of the interface and the layout;
- In the wrap-up/evaluation meeting (*D*), the group decided on which prototype to choose and what functions to be eliminated from the prototype.

At the beginning of each session, the experimenter introduced the browser interface to the subject. The subjects were then free to browse through one pre-selected meeting recording (which is not used in the real experiment). They could take as much time as they needed to familiarize themselves with the interface.

The main task is as follows:
"*In 45 minutes or less, write a report to summarize the decisions made in the four meetings for upper management*.

We choose 45 minutes as the time constraint of this task is because: Users are able to summarize a decision in depth by browsing through the summaries of a series of four meetings in 45 minutes Murray et al. (2006), and we want to know how users behave when asked to do decision debriefing (i.e., summarize all decisions made) in the same series of meetings given the time constraint. Also, if the decision debriefing task is indeed a task that can be completed in 45 minutes, will users do equally well when presented with either a generic summary or a decision-focused summary?

Because some subjects in the pilot study expressed the need to be reminded of the time remaining in the experiment, the experimenter signalled the subjects twice before the end of the experiment, once at 25 minutes and again at 40 minutes into the experiment. The subjects could also signal the experimenter to end the session if they finished the task early. During the session, all the user behaviors were recorded in log files, and the user-generated decision minutes were logged separately.

At the end of each session, the experimenter asked the subjects to explain how they used the browser interface to find out about the decisions made in the meetings. Subjects were also asked to fill in a post-questionnaire (c.f. Appendix D) about their perceived task success.

7.3.2 Meeting Corpus and Gold Standard

Among the many natural meetings that have been recorded in the context of the ICSI Meeting project (Janin et al., 2003), the CALO project (CALO, 2006) and the AMI project (Carletta et al., 2006), we chose the AMI meeting corpus for our work because the participants in this corpus are making decisions as a group in a series of meetings that imitates a typical product design cycle, starting from a kick-off meeting and ending with an evaluation meeting.

7.3.2.1 Decision point annotation

Recall from Chapter 3.2.3.1 that annotators were asked to navigate the recordings of one series of four meetings and to summarize the decisions made in these meetings into a list of "decision points" (as shown in Figure 7.2). (For a complete list of these bullet point-listed summaries, please refer to Appendix E.)

In this phase of the annotation procedure, the set of decision points that were noted by *two ore more* annotators are used as the gold standard set of decision points. In the meeting series used in this study, the meeting participants reached 6, 10, 8, and 6 decisions respectively.

7.3.2.2 Decision-focused extract annotation

Then, another group of three annotators were asked to go through the dialogue acts in each meeting one by one, and judge if they could be annotated as a "decision-related dialogue act (decision DA)", i.e., if they supported any of the decision points. If so, the DA is "linked" to the decision point.

Independent Variable (IV)	Levels
Automatic Extract Type	General/ Decision-focused
Decision-focused DA Selection Type	Manual DDA/ Auto DDA
Transcription Type	Manual/ ASR

Table 7.1: *Experimental design of the task-oriented evaluation: independent variables (IV).*

DV	Factors
Task effectiveness (Objective)	Task difficulty, completeness, effort required, Proportion of clicked extracts, reading speed, Productivity, use of media and summary content (See Table 7.4)
Perceived success (Subjective)	Ease of use, task completeness, Decision coverage and comprehension (See Table 7.5)
Report quality (Objective)	Overall quality, completeness, conciseness, trustworthiness, style

Table 7.2: *Experimental design of the task-oriented evaluation: dependent variables (DV).*

After each annotator finished their annotations, the ground truth "decision-focused extract" of each meeting (e.g., those used in the decision-focused summary display in Figure 7.1) is then generated by collecting the set of decision DAs that were extracted by *one or more* annotators. For detailed description of the decision-related DA annotation, please refer to Section 3.2.3.1. The level of inter-coder agreement is shown in Table 3.9.

7.3.3 Meeting Browser Interface

The meeting browser (cf. Fig. 7.1) used in this evaluation is an enhanced version of the J-Ferret meeting browser (Wellner et al., 2004), which is designed to present the meeting recordings augmented with additional types of information to aid users in browsing

(Carletta et al., 2006). The enhanced version consists of three basic components: the audio-visual recording playback facility (top), the transcript display (lower left), and the extractive summary display (lower right).

Each subject is equipped with a headphone so that they can listen to the audio recordings whenever necessary. Users can play the audio-video recording from the beginning. Users who are interested in a particular decision DA can click on that "***content button***" in the display and be led to the point where the DA was uttered in the dialogue. Each of the decision DA buttons in the summary display is time synchronized to the location of the decision DA in the audio-visual recording as well as in the transcript.

There are five tabs on the top of the browsing interface: (1) the first four tabs take users to each of the four meetings in the series chosen for display, and (2) the last tab is the "writing tab", where users are asked to type in their summaries. Users can switch between these tabs at will. During the experiment, a logging tool in the back-end records all the clicking and typing behaviors. With this log, we can analyze the use of the different components in the browser, such as the summary display and the audio-video playback facility, as well as the report typing behavior, e.g., how many characters were deleted, inserted, and substituted by each subject.

7.3.4 Manual decision-focused extracts

Because we do not want to assume that we will always have multiple annotators for producing the decision-focused extracts, in the following experiment, we chose one set of decision DA annotations from a most experienced annotator to be displayed to the users. Since this set of DAs was selected by only one annotator, it would have missed some decision points listed in the gold standard abstractive summaries. Therefore, we also analyzed the proportion of the decision points that can be captured by the set of annotations we chose: In the two meetings we have gold standards, i.e., the decision that have been selected by multiple annotators, the chosen set captured 88% (14 out of 16 decision points).

7.3.5 Automatic decision-focused extracts

As ultimately we wish to generate decision-focused summaries automatically, we also evaluated the effectiveness of the automatic decision-focused extracts. The automatic decision-focused extracts used in this experiment were generated by the state-of-the-

art decision detection algorithm developed in Chapter 5, which trained MaxEnt models to classify whether each DA is decision-related or not and constructed ensembles from multiple MaxEnt models that were trained with different knowledge sources. The knowledge sources used to generate the models used in this evaluation include the topic of the current discussion and words spoken, pitch and energy level, preceding and following pauses, and annotations of dialogue act types. These features have been shown to be characteristics of the contexts surrounding the decision-related DAs in the analyses in Chapter 4. The intrinsic evaluation showed that this model yields 30-40% of inconsistencies with the ground truth data.

Note that the same number of decision DAs as that in the ground truth data were then selected to form an automatic decision-focused extract. This is to ensure the length of the automatic extracts would be the same as the manual ones so that no effect of summary length would be introduced in the evaluation.

To understand how many decision points would be missed by using the automatic decision-focused extracts, we navigated through the set of automatic decision DAs that were displayed with the transcripts and recordings in the browser and recorded the number of decision points in the ground truth data that could not be found with the hints from the automatic DAs. For example, among the eight decisions contained in the ground truth data of Meeting *C*, this method would miss 3-6 major decisions; A more casual reading of the transcripts by the human judge would result in the misses of two more decisions and the misinterpretation of one decision. Among the 30 decisions found in the ground truth data of the series of meetings used in this experiment, the automatic DAs could be used to identify 17-23 of them (57-76%), depending on how carefully would the users read through the neighboring transcripts.

7.3.6 Automatic general-purpose extracts

We leveraged the automatic extractive summaries generated in Murray (2007). The summaries were trained with a feature-based approach that combines prosody, meeting structure, speaker status, and lexical informativeness. The level of lexical informativeness is determined by the speaker-dependent inversed document frequency (su.idf) measure, which extends tf.idf to capture the speaker-specific informativeness.

The idea is as follows: If a word that has been rarely seen in a speaker's previous speech has suddenly become prominent in his speech, it is a good indicator that the word is more important to be summarized. In Equation 7.3, $N(t)$ denotes the number of

times the term t occurs in the given document, T the total number of term occurrences in the given document, D_t the number of documents containing the term t, and D the total number of documents.

$$tf(t,d) = \frac{N(t)}{\sum_{k=1}^{T} N(k)}, \tag{7.1}$$

$$idf(t) = (-log(\frac{D_t}{D})), \tag{7.2}$$

$$tf.idf = tf * idf \tag{7.3}$$

In Equation 7.5, $N(r)$ denotes the number of words that have been spoken by speaker r, S the total number of speakers, D_t the number of documents containing the term t, and D the total number of documents.

$$su(w) = \frac{1}{S}\sum_{s}(-log(\frac{\sum_{s' \neq s} tf(w,s')}{\sum_{r \neq s} N(r)})), \tag{7.4}$$

$$su.idf = su(w) * \frac{s(w)}{S} * idf \tag{7.5}$$

The lengths of the general summaries represent compressions of approximately 40%, 32%, 32% and 30% of the four meetings in the series respectively.

7.3.7 Automatic speech recognized transcription

The ASR transcription used in this experiment was generated with a state-of-the-art program, which on average recognizes words with 30%-40% error rate. Section 3.1.2 contains a short summary of the techniques used. More technical details can be found in Hain et al. (2005). The ASR results rendered the interpretation of some decisions difficult. A few examples are demonstrated in Table 7.3.

7.3.8 Experiment design

Our research questions are concerned with the effect of automatic summary type on users' task performance and perceived success. Our hypothesis is that a more succinct, query-focused summary (a small portion of the general purpose extractive summary) would help users prepare a meeting minute for upper management more effectively and feel more confident in the meeting preparation work. In addition, we also test the

	Manual transcript	ASR output
Meeting *B*	”So you never have to change the battery”	”So you never have to change that ”
Meeting *C*	”the buttons can be fruit-shaped”	”the buttons can be fair share”
Meeting *C*	”the ideal of wheel like ipod”	”the idea of oh we all like five um”
Meeting *C*	”and the cool case”	”in the cold case”
Meeting *C*	”titanium”	”taking”
Meeting *C*	”a fashion”	”old-fashioned”
Meeting *D*	”we’re having at least three colours”	”they were really street however”

Table 7.3: *Possible misinterpretation of decisions resulted by ASR outputs.*

impact of automation on the performance of the decision-focused summary. The subjects we recruited were randomly assigned into four groups. Each group was asked to accomplish the decision debriefing task, using one of the following summary displays embedded in the meeting browser:

- BASELINE (AE-ASR): automatic general-purpose extracts (AE)[3], automatic speech recognized (ASR) transcription. The reason that we chose the extractive summaries as baseline is mentioned in Section 7.3.1: we want to be able to compare the effect of a generic summary and a query-focused summary on user performance. In addition, the quality of the extractive summary produced by Murray (2007) is the state-of-the-art. It has been proven as useful for finding information to realize a decision in depth and in turn served as a high bar for the decision detection model to beat.

  There are certainly many other possible ways to create baseline, in particular, randomly extracting DAs into the summary. We did not consider these other ways due to the belief that the generic summary is the hardest-to-beat baseline we can possibly obtain. Also, due to the decision sparseness in the nature of meeting behaviors,the chance of seeing decision-critical information in the 1% of randomly chosen DAs is going to be minimal.

- AD-ASR: automatic decision-focused extracts (AD)[4] , ASR transcription.

[3]We used what is generated in Murray (2007). The lengths of the general summaries represent compressions of approximately 40%, 32%, 32% and 30% of the original meeting length respectively.

[4]The automatic decision-focused extracts used in this experiment were generated by our state-of-the-art decision detection program reported in Chapter 5, which predicts decision DAs with 60%-70% agreement with the model summaries in the series of meetings used in this experiment.

The labels used in the AD-ASR condition is generated by projecting the position of the automatically identified decision DAs on the ASR transcripts. In Section 5.6.2 we have proven that using ASR or REF transcripts does not yield significant differences in meeting decision detection. Therefore here we assume the same set of automatically extracted DAs for both the AD-ASR and AD-REF conditions. Both the transcripts and the decision-focused extracts are hence based on the error-prone ASR outputs, leading to the occasional misinterpretation of what the meeting participants were saying in the recordings.

- AD-REF: automatic decision-focused extracts[5], manual transcription (REF).
- TOPLINE (MD-REF): manual decision-focused extracts (MD), manual transcription (REF).

We designed our experiment to test the hypothesis that a decision-focused summary display benefits users more than a general-purpose display for accomplishing the decision debriefing task. There are many techinques commonly used for evaluating summarization systems. On the one hand, we can gauge the summary quality with the intrinsic measures of summary "readability" (e.g., conciseness, understandability, grammar, and style) (Minel et al., 1997; Saggion and Lapalme, 2000) and "informativeness" (e.g., how much information is preserved in the summary). Mani et al. (1999) have found an inverse relation between the two dimensions. In the evaluation of this study, we will evaluate on both, but focus more on the latter one.

On the other hand, we can evaluate the use of the summaries with real users. Tasks that have been proposed previously for evaluating meeting browsers as a whole. For example, Wellner et al. (2005) have proposed the Browser Evaluation Test (BET) to evaluate whether users can identify most of the observations of interest (e.g., who is wearing red, what movies have been recommended by one of the meeting participants) from the meeting archives using the browser in test. Different from previous user studies that focused on user satisfaction, this BET set provides a less costly way to identify observations of interest such that an objective measure of browser effectiveness can be automatically computed based on user performance. Popescu-Belis et al. (2008) further experimented with a task-based BET (TBET) set to examine the effectiveness (precision) and efficiency (speed) of different meeting browser designs.

[5]AD-REF uses the same version of automatic decision-focused extracts (AD) applied in AD-ASR. The only difference is that this set of ADs are mapped onto manual transcription under the AD-REF condition.

In this study, we also test our hypothesis in a task-oriented environment. However, we do not assume the BET setting which evaluates how well a user can answer factual questions by observing generic observations of interest from the archives. This is because we want to concentrate on browser designs that are effective in real-life organizational settings, and in such setting, capturing all possibly important facts of a lengthy archive is less important to cutting to the point.

Although we do not adopt these evaluation frameworks, our evaluation plan still takes into account all the previously proposed intrinsic and extrinsic metrics. The independent and dependent variables are shown in Table 7.1 and Table 7.2. The independent variables are tested between subjects.

The dependent variables are classified into three categories:

1. ***Task effectiveness***: First of all, the user-generated decision minutes are manually evaluated against the gold standard decision points (which were obtained using the procedure in Section 7.3.2.1). If a gold standard decision point is found in the user-generated minute, a "decision hit" is marked. Task effectiveness is measured by the coverage of the decision hits, that is, the percentage of the gold standard decision points that have been correctly listed in each of the user-generated decision minutes.[6]

2. ***Report quality***: Several aspects of the user-generated decision minute quality are rated on a 1-7 Likert scale (1 = most positive, and 7 = most negative)[7]. These aspects include the overall quality, completeness, conciseness, trustworthiness[8] and writing style.

   One human judge was used in the intrinsic evaluation of report quality. This judge was blind to the condition used to generate the summaries. The human judge used a set of criteria to determine the score for each of these aspects. For example, the level of conciseness is determined by the following criteria: (1) When the summary is mainly consisted of keywords; (2) When each decision is depicted in one sentence; (3) When each decision is depicted in one or

[6]We followed the tradition of the summarization research community to use recall-based measures, e.g., ROUGE Lin (2004). Also, the subjects assumed different writing styles in the decision debriefing task. Some of them wrote very thorough descriptions of the decisions they found, while some of the others described decisions in keywords. The heterogeneous styles have rendered the calculation of precision impractical.

[7]The rating of each aspect is determined by an annotator, following a guideline developed in the pilot study.

[8]The level of trustworthiness reflects how useful a meeting archive user who did not attend the meeting will find the summary to be for meeting preparation.

more sentences; (4) When multiple decisions are depicted in one paragraph; (5) When the author includes some decision-irrelevant information (e.g., those that would appear in general-purpose summaries but not decision-focused ones) in the summary; (6) When the author fills half the summary with information that is decision-irrelevant; (7) When the author include mainly decision-irrelevant information in the summary.

As there can still be indiviudal differences for determining the scores, we used only one judge in this evaluation to ensure that the assessment of scores would be consistent across all summaries.

3. ***User perceived success***: Finally, Table 7.5 lists the self-reported measures of the level of perceived success and usability, reported on 5-point Likert-scales[9] in the post-questionnaire.

In addition to the three evaluation criteria, we also developed a number of quantifiable measures to understand user behavior when given the different types of summary displays. These measures, which are computed from the log files, include task completeness, proportion of clicked extracts, reading speed, productivity, usage of media, and usage of extracts to correct summary writing.

For example, the proportion of clicked content buttons in extracts is measured by counting the number of "***content clicks***"[10]. A more detailed account of the log-based measures of user behaviors is presented in Table 7.4.

7.4 Results

In the context of the decision debriefing task, this study aims to answer the main question:

- Whether general-purpose extractive summaries–which extract generically important dialogue acts that reflect overall meeting content–could be improved by focusing on only the decision-related dialogue acts.

In addition, because we would like to know how much automation degrades the usefulness of the summary display, we also address the following two questions:

[9]We did not use the 7-point scales in the questionnaire, since the questionnaire is a modified version of that used in Murray (2007) and we would like to understand if the level of perceived success by the users of general extractive summaries and decision-focused summaries are in the same ball park.

[10]The set of decision-related content buttons (each reflecting one decision-related DA in the extractive decision-focused summary) that have been clicked by the user.

Factors	Measures
Task completeness	Number of meetings summarized in the minute
Time to write first decision	Time to type first character
Proportion of clicked extracts buttons in extracts	Number of content clicks, normalized by total number of content buttons
Writing speed	Normalized frequency of switching to the writing tab
Reading speed (Extract)	Normalized frequency of content clicks
Productivity (by writing time-stamp)	Average frequency of insertions
Productivity (by report word count)	Number of words in user's report (edited)
Usage of media	Number of times user played the audio or video
Usage of extracts to correct writing	Number of content clicks in the preceding 2 minutes of a writing tab click

Table 7.4: *Log file-based measures of task effectiveness.*

- Can the automatic decision-focused extracts help users achieve performance comparable to that obtained by navigating the manual extracts?
- Does operating on the transcription produced by automatic speech recognition (ASR) as opposed to manual transcriptions affect user performance significantly?

In the following section, we first perform ANOVA test to examine the overall main effect of the conditions on the dependent variables, i.e., the task effectiveness, perceived success, report quality. We then perform Tukey HSD tests to conduct multiple comparisons across each pair of conditions. In particular, we focus on three pairs of conditions: AE v.s. AD, AD v.s. MD, and ASR v.s. REF.

7.4.1 Main Effect of Summary Display Type on Decision Debriefing

In this section, we report the results of task effectiveness and report quality obtained from the analysis of the log files and the minutes. We also assess the user's behavior in the use of the different types of summary display.

7.4.1.1 Task effectiveness analysis

The data analysis shows that, on average, users who were given the decision-focused summary display yield more decision hits than the general-purpose one (Figure 7.3).

DV Factors	Post-questionnaire Statement
Perceived ease of use	Q1: I found the meeting browser intuitive and easy to use.
Perceived ease of search	Q2: I was able to find all of the information we needed.
Perceived efficiency	Q3: I was able to efficiently find the relevant information.
Perceived task completeness	Q4: I feel that we completed the task in its entirety.
Perceived comprehension (general)	Q5: I understood the overall content of the meeting discussion.
Perceived task success (decision)	Q6: I was able to efficiently find the decisions.
Perceived task difficulty	Q7: The task required a great deal of effort.
Perceived pressure	Q8: we had to work under pressure.
Perceived system usefulness	Q9: I had the tools necessary to complete the task efficiently.
Perceived lack of support	Q10: I would have liked additional information about the meetings.

Table 7.5: *Questionnaire-based measures of user perceived success and usability.*

To determine whether the differences are statistically significant, an analysis of variance was performed. The meeting summary display type was found to have a significant main effect on task effectiveness ($F(3,31) = 13.832;\ p < 0.001$). The best performing subject was able to use the TOPLINE MD-REF (manual decision extract, manual transcription) browser to find 27 out of all 30 decision points (90%).

Because we noticed that some of the subjects were not able to finish the last meeting, the decision hits are also measured in the first three meetings to investigate the effect of experimental conditions on task completeness. The benefit of using the decision-focused summaries holds true when the decision hits are measured in the first three meetings only, indicating that there is not a major impact of the conditions on task completeness in the context of this experiment design.

7.4.1.2 Report quality analysis

A condition (4) x quality (5) analysis of variance on the decision minute ratings (Table 7.7) shows that the meeting summary display type also has a significant main effect on the reported overall quality ($F(3,31) = 3.324;\ p < 0.05$). With the more precise information in the decision-focused summary display, the subjects are able to generate decision minutes of higher quality.

	TOPLINE	AD-REF	AD-ASR	BASELINE
Task completeness	4	4	4	4
Time to write first decision	6.3	2.64	1.38	4.08
Proportion of clicked extracts	0.53	0.61	0.51	0.12
Writing speed	1.97	1.84	2.19	1.40
Reading speed (Extract)	1.90	1.99	1.70	2.65
Productivity (frequency)	0.56	0.52	0.53	0.57
Productivity (characters)	2,452.3	2,013.43	1,641.56	1,758.67
Usage of media	23.60	17.71	33.44	15.56
Usage of extracts to correct writing	9.6	9.6	7.2	2.4

Table 7.6: *Task effectiveness measures based on user-clicking behavior.*

Although we did not find any significant effect on other measures of report quality, a finer-grained comparison shows that there are differences between the Baseline condition and the other three conditions in terms of conciseness and trustworthiness. The general purpose display performs fundamentally worse than the decision-focused on in these two aspects. In addition, the ASR transcripts are detrimental to the completeness and writing style.

Criterion (1-7)	TOPLINE	AD-REF	AD-ASR	BASELINE
Overall Quality	2.5	2.4	3.6	3.9
Completeness	3.1	2.9	3.8	3.4
Conciseness	2.4	2.7	2.6	3.4
Writing Style	2.6	2.1	3.3	3.4
Trustworthiness	1.9	2.0	1.8	2.4

Table 7.7: *Quality assessment of the subjects' minutes. Results are obtained on a 7-point scale: the lower the score, the better the minute quality.*

7.4.1.3 Perceived success analysis

The average ratings reported in the post-questionnaires (cf. Table 7.8) suggest that the decision-focused display is perceived to be easier to use ($F(3,31) = 4.819$; $p < 0.05$)) and require less effort ($F(3,31) = 4.343$; $p < 0.05$). The subjects using the decision-focused display also find themselves able to retrieve the relevant information more

efficiently ($F(3,31) = 8.710$; $p < 0.01$), and absorb the decisions made in the meetings more effectively ($F(3,31) = 4.714$; $p < 0.05$).

Criterion (1-5)	TOPLINE	AD-REF	AD-ASR	BASELINE
Perceived ease of use (interface)	4.4	4.1	4.3	3.6
Perceived efficiency	3.9	3.4	3.6	3.3
Perceived comprehension (general)	4.6	4.6	4.1	4.1
Perceived task success (decision)	4.3	4.3	3.8	3.7
Perceived task difficulty	2.6	2.9	2.9	3.7
Perceived pressure	2.8	3.8	2.7	3.4
Perceived system usefulness	4.4	4.3	4.1	.4.1

Table 7.8: *User perceived task success. Results are obtained on a 5-point scale (5 = agree strongly to 1 = disagree strongly).*

7.4.1.4 User behavior analysis

An ANOVA test on the log-based data reveals that, compared to the subjects in the baseline condition (automatic general-purpose extract, ASR transcription), subjects in the rest of conditions click on a significantly higher proportion of the extracted decision DAs to write minutes ($F(3,31) = 9.878$; $p < 0.001$) and rely more on the extract contents to modify their minutes ($F(3,31) = 21.715$; $p < 0.001$). (See Table 7.10.)

7.4.2 Pairwise Comparison

Having examined the main effects across the four conditions, we now investigate whether there exist significant differences between each pair of conditions. As explained in Section 7.3.1 and 7.3.8, each of these pairs of conditions captures an important distinction in summary display designs.

Table 7.9 shows the Tukey HSD test results for all the different pairs of conditions. Each cell represents a significance code of the test: "*" for $p <= 0.05$, "**" for $p <= 0.01$, "***" for $p <= 0.001$.

7.4.2.1 Decision-focused Extracts v.s. General-purpose Extracts

The first pair we compare is the AD-ASR and Baseline (AE-ASR) conditions. Given that the decision-focused summaries are more effective than the general-purpose sum-

DV Measure	Cond0 C̃ond1	Cond0 C̃ond2	Cond0 C̃ond2	Cond1 C̃ond2	Cond1 C̃ond3	Cond2 C̃ond3
Decision hits (all 4 meetings)	NS	**	***	NS	**	NS
Decision hits (first 3)	NS	**	***	NS	**	NS
Proportion of clicked extracts	NS	NS	***	NS	***	**
Use of extracts to correct writing	***	***	***	NS	NS	NS
Usage of media	23.60	17.71	33.44	15.56		
Usage of extracts to correct writing	9.6	9.6	7.2	2.4		

Table 7.9: *Tukey HSD test results, with NS denoting "not significant". Cond 0: AE-ASR; Cond 1: AD-ASR; Cond 2: AD-REF; Cond 3: AD-REF.*

maries for the decision debriefing task, we wish to determine whether the effectiveness remains when a more error-prone automatically generated summary is used in the interface. To determine patterns that were not specified a priori, posteriori pairwise comparisons were performed.

First, we examined the decision hits across the conditions that use the automatic general-purpose display (BASELINE) and the one that used the automatic decision-focused display (AD-ASR). The percentage of decision hits (as reported in Figure 7.3) shows that focusing on only the decision-related information results in greater task effectiveness – on average, increasing the number of decision hits over that yielded with the general-purpose display by 36%. Moreover, the decision minutes generated by the subjects who use decision-focused summaries tend to exhibit better overall quality and conciseness.

Further analysis of user behaviors reveal that, even when the summary contains ASR errors, the subjects still rely more on the decision-focused display to summarize meeting decisions. The decision-focused display is found to significantly increase the proportion of the content buttons (each representing one extracted dialogue act) that has been clicked by the subjects during the decision debriefing task ($p < 0.01$; Tukey's test).

7.4.2.2 Automatically Generated Extracts v.s. Manual Extracts

The question that emerges naturally next is how much performance degradation results from replacing the manual summary with its automatic version (which contains 30%-

	TOPLINE	AD-REF	AD-ASR	BASELINE
Proportion of clicked extracts	0.53	0.61	0.51	0.12
Usage of media	23.60	17.71	33.44	15.56
Usage of extracts to correct writing	6.84	1.54	0.93	0.66

Table 7.10: *Task effectiveness measures based on user behavioral cues.*

40% inconsistencies with the ground truth). The answer would provide useful guidance for the design of meeting browsers, and may provide support for the development of automatic machinery for query-focused speech summarization.

Therefore, we compared task effectiveness (in terms of the number of decision hits) and report quality between the condition that uses manual decision-focused summaries (TOPLINE) and the one that uses automatically generated summaries (AD-REF). Although the overall quality of minutes in the two conditions does not differ significantly, the automatic extractive summaries yield on average three fewer decision hits (21%).

To further understand whether the errors in the summary resulted in any systematic difference in user behavior, we examine the log files (cf. Table 7.10). We expected that users would prefer to use the meeting summary display to find decisions when the summaries are reflective of the actual decisions made. The post-hoc test results match the expectation: Using the automatic version of the summary (AD-REF) instead of the manual version (TOPLINE) significantly decreased the use of the summary display prior to writing correction (Tukey's test, $p < 0.001$).

However, this difference in task effectiveness and user behavior does not seem to affect the subjects' perceived success towards the task and ability to produce quality minutes: no significant difference was found in any of the subjects' ratings in the post-questionnaire and the minute quality ratings for the two conditions.

7.4.2.3 Effect of Transcription Type

Because our ultimate goal is to design a meeting browser that can be used as soon as a meeting ends, it is important to study whether operating the browser on error-prone automatic speech recognition (ASR) transcription (which contains 30%-40% errors) affects task effectiveness and report quality.

To examine the performance degradation caused by the ASR transcript display, post-hoc tests were also performed across the conditions that operate on ASR tran-

scription (AD-ASR) and on manual transcription (AD-REF). The assessment results of report quality (cf. Bar AD-REF and AD-ASR in Fig. 7.3) suggest that displaying decision-focused summaries on manual transcripts helps the subjects find 39% more (on average, 4 to 5) decision hits than displaying the summaries based on ASR transcripts. The advantage is also confirmed in the post-hoc test ($p < 0.01$).

Further analysis of the decision minute quality (cf.Table 7.7) shows that users who browse summaries on manual transcripts are likely to produce decision minutes of better overall quality and completeness. In addition, the more readable transcripts allow the subjects to allocate more of their time to absorbing relevant information, rather than understanding meeting content. In turn, the decision minutes generated by this group of users can be better appreciated by readers.

Examination of user behaviors (cf. AD-REF and AD-ASR in Table 7.10) also shows the transcription type to have effects on the usage of the summary display for writing decision minutes ($p < 0.05$). The less helpful displays increase the level of perceived pressure ($p < 0.05$) reported by the subjects (cf. Row 6 in Table 7.8). Compared to the errors introduced by the automatic decision-focused extracts, the errors introduced by the ASR transcripts have a greater negative impacts on the users' performance level.

7.5 Discussion

DV	Measures
Task Effectiveness	Proportion of clicked extracts $(F(3,31) = 9.878; p < 0.001)$ Usage of extracts before writing $(F(3,31) = 21.715; p < 0.001)$
Report Quality	Overall quality $(F(3,31) = 3.324; p < 0.05)$
User Perception	Easy to use $(F(3,31) = 4.819; p < 0.05)$ Effectively finding information $(F(3,31) = 8.710; p < 0.01)$ Required effort $(F(3,31) = 4.343; p < 0.05)$

Table 7.11: *ANOVA results of task effectiveness for subjects across all four conditions.*

The results of this study verify our experimental hypothesis. Displaying decision-focused summaries in the meeting browser helps users to obtain an overview of the decisions from multiple meeting recordings more effectively than general-purpose summaries. The decision-focused summary, obtained by filtering out the dialogue acts irrelevant to decisions, was found to improve not only task effectiveness, but also the overall quality of the subjects' minutes. The users in the focused summary conditions read through a higher proportion of summary material to find relevant information and relied more on summaries to prepare and correct the decision minutes they wrote.

It also help increase user-perceived efficiency. However, as the decision debriefing task in our experimental design is considered as a straightforward task to be completed within the time allowed, we cannot draw conclusions about the impact of the decision-focused extracts on efficiency.

Having established the advantage of the decision-focused summary, our investigation further examined the impact of (1) using automatically generated decision-focused summaries and (2) that of using ASR transcripts to produce summaries and display in the transcript plug-in.

The first examination showed that, participants who read the automatic decision-focused summaries outperformed those who read the general-purpose summaries in the decision debriefing task. The automatic summary users clicked on more extracted dialogue acts to understand the content and correctly identified more decisions ($r = 3.573$, $p < 0.001$).

Although the automatic summary users did not find as many decision points as those using manual summaries, they were able to produce decision minutes of similar quality. One explanation for this could be that the automatic summaries correctly identified parts of the decision points, and the automatic summary users could leverage the parts of the automatic summaries to find relevant information for summarization. hence yielding minutes of these decision points at similar quality. However, as some of the decision points were not captured in the automatic summaries in the first place, the automatic summary users were likely to miss these decision points in their minutes.

7.5.1 Use of audio-video aids

From the post-experiment debriefings, we observed two main strategies adopted by users to find the decision points that were not clearly presented in the extractive summaries: (1) Some users attempted to go through the extracted DAs in the summary

display one by one looking for relevant information in the surrounding context in the transcript; (2) Others turned to the audio-video recordings to find the missing decisions. The two coping strategies can be distinguished by their usage of media. Table 7.12 presents the proportion of subjects that have high and low usage of the audio video aids.

It appears that when the manual transcripts are in the display, e.g., in the TOPLINE and AD-REF conditions, the choice of strategy was based on individual differences, and a majority of the users preferred to use the decision-focused summaries rather than the audio-video aids. However, when the error-prone ASR transcripts are in the display, e.g., in the AD-ASR and BASELINE conditions, the choice of strategy was noticeably affected by the the type of summary display. Comparing Columns AD-ASR and BASELINE in Table 7.12 illustrates that the AD-ASR users tended to make more usage of the audio-video recordings. This is because the ASR transcripts are difficult to understand by themselves, and it is therefore important to find additional hints from the summary display; However, as the summaries presented in the AD-ASR display are often short and error-prone, the audio-video recordings are necessary for accomplishing this task.

Media Usage	**TOPLINE**	**AD-REF**	**AD-ASR**	**BASELINE**
Low (<30)	70.0%	85.7%	44.4%	88.9%
High($>=30$)	30.0%	14.3%	55.6%	11.1%

Table 7.12: *The proportion of subjects who had low and high usage of audio-video recordings: Low=playing recordings less than 30 times; High=playing recordings greater than or equal to 30 times.*

Figure 7.4 demonstrates the effect of audio-video usage on task effectiveness and user-perceived success. The analysis reveals that the AD-ASR and BASELINE users who turned to the audio-video browsing strategy (i.e., those with high usage of audio-video aids) were more likely to miss decisions in the archives. Interestingly, the lower task success rates did not affect the ratings of user-perceived success. For example, the group of high media usage users under the AD-ASR condition, who on average yielded lower task effectiveness, still perceived a high level of task success. The finding coincides with the subjects' comments that, although the audio-video recordings are difficult to use, they provided grounds for decision understanding.

7.5.2 Use of decision-focused extracts directly on decision debriefing

This experiment is designed with the hypothesis that the users need to leverage the aids (i.e., transcripts, audio-video recordings) from the browser to achieve the decision debriefing task. However, is it possible to use the decision DAs directly to obtain an overview of decisions without having to go through the process of producing abstractive summaries? To answer this question, we further evaluated the task effectiveness of the decision DAs on the task of interpreting the decision points in the series of meetings used in this experiment. Using the decision-focused extracts (as shown in Figure 7.6; more in Appendix F) directly can help interpret 76% of the decision points. Following is a breakdown of the decisions that were missed or misinterpreted in the decision-focused extracts.

- Meeting *A* (6 decisions): two major decisions were missing in the extracts.
- Meeting *B* (10 decisions): one major decision was difficult to interpret from the text selected in the extracted DAs (i.e., "B*ut you could maybe have it in a little charging station like a mobile phone, or like a little cradle for your iPod . So you don't ha you got like a rechargeable battery.*").
- Meeting *C* (8 decisions): one major decision was missing in the extracts, and one was difficult to interpret from the text (i.e., "*Let's go with a simple chip?*")
- Meeting *D* (6 decisions): two major decisions were difficult to interpret from the texts, one due to the missing context (i.e., "*I think I'm leaning towards the potato*.") and another due to the missing connections between the extracted DAs (i.e., no connections was shown between "*but um the um feature that we considered for it not getting lost*" and "*Um we so we do we've decided not to worry about that for now*").

However the level of understanding achieved with solely the decision-focused extracts is shallow. When the goal is to obtain in-depth understanding of the contexts surrounding the decisions, displaying the extracts with the transcripts and audio-video recordings in the browser is still necessary.

7.6 Conclusion

This study has verified our experimental hypothesis: Existing meeting summarization systems, which provide a general-purpose summary display, can be improved by refo-

cusing the summaries with regard to user's information need. For users who require a quick overview of decisions, the decision-focused summary display was found to improve not only the actual task effectiveness, but also the overall report quality. The browser interfaces that come with the decision-focused summary display are also rated as easier to use.

Users also perceived the decision-focused summaries useful in helping them to find all relevant information and understand the decisions more efficiently. However, due to the experiment setting we could not draw conclusions on the actual improvement of decision-focused summaries on efficiency.

In addition, we evaluated the impacts of automation on the decision debriefing task. The findings are as follows: (1) The automatically generated decision-focused summaries, which contain 30%-40% inconsistencies with the gold standard manual summaries (a, still assist users in producing high quality decision minutes and feeling confident about their performance on the decision debriefing task, despite some reduction in decision hits. (2) The ASR transcription has a greater negative impact on the actual task effectiveness and the quality of minutes. Another side effect of the ASR display is an increase in the level of user-perceived pressure.

Further investigation demonstrates a correlation between task effectiveness and usage of the summary display. As the content in the decision-focused summary is more closely tied to the user needs, participants who use these summaries (as opposed to the general purpose ones) rely more on the summary to find relevant information and, in turn, achieve higher performance.

Finally, the examination of user's media usage and coping strategies suggest that there exists an individual difference in the user's preference of whether to use the summary display or the audio-video playback facility to find relevant information. However, when the decision-focused summary is displayed with the ASR transcripts, users are often forced to view the audio-video recordings, since it is difficult to use the other two displays (i.e., transcript, audio-video facility).

The finding is similar to that in Whittaker et al. (1999) that when scanning and extracting information from speech archives (such as broadcast news and voice mails), direct access to the original speech is still necessary for some users. First, the ASR errors prevent accurate intepretations of the original speech. Second, intonations are also essential for interpretations. This is even more true when the input sentences are not complete paragraphs, but individual sentences picked from dialogues.

This also suggests that there is a need to provide additional interface assistance to

facilitate this group of users when ASR transcripts are used and may affect the comprehensibility of a succinct decision-focused summary. Possible interface enhancements include a switching device that allows users to freely go from the view of a decision-focused summary back to that of a general-purpose summary.

In short, this evaluation provides the first account of how query-focused summaries facilitate in users' understanding of multimedia archives. However, there are concerns about whether our approach is successful only when used to summarize well-structured meetings such as those product design meetings recorded in the AMI corpora. These concerns have first been addressed in the way the meetings were recorded – the product design meeting scenario was designed to be as close to what might be happening in real life. In addition, in our observation, the meeting participants do not always summarize decisions in the end. an empirical analysis of the positions of the decision DAs in the meeting recordings has confirmed this: only 26% of the decision-DAs occurs in the last five minutes of the meetings (which last 30-40 minutes on average), and the average position of the decision DAs are in the middle part of the meetings (around 60% into the meetings). Also, since the subjects of this evaluation were not provided with any material relevant to the meeting scenario other than those that came with the meeting browser, this evaluation is expected to reflect how the users will use the focused summary-enhanced browser to gather information needed for debriefing query-related information.

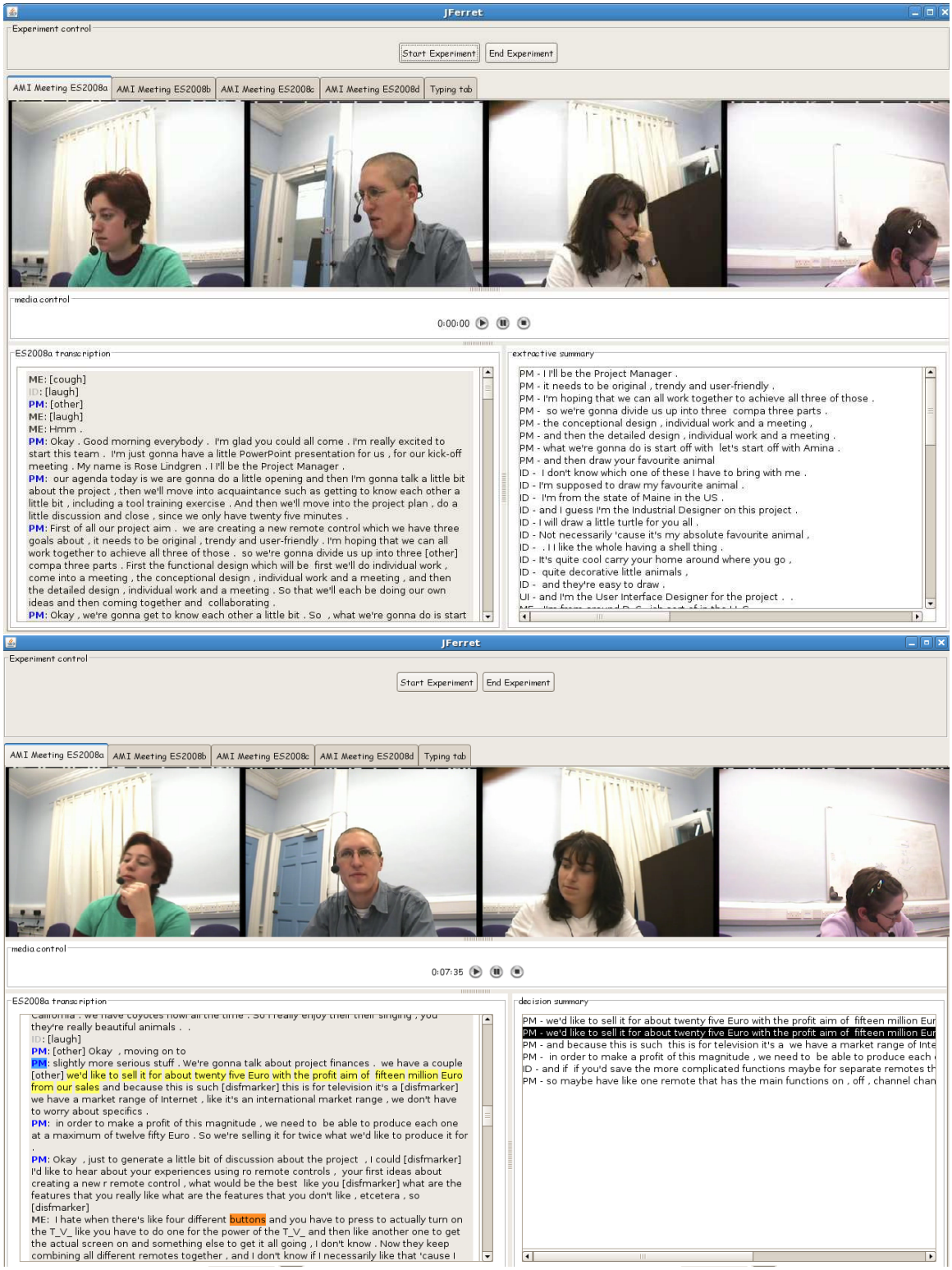

Figure 7.1: *Example AMI browsers. Each browse is composed of three components: the playback facility of audio-video recordings (top), the transcription (lower left), and the extractive summary display (lower right). The two browsers differ in the summary presented in the display: the browser on the left presents the general-purpose summaries, and the one on the right presents the decision-focused summaries.*

The group decided not to define the target user group by a specific age range but simply by interest in fashion and simplicity. The remote will feature a locator function and large buttons. The remote will incorporate both simple and complicated functions, hiding the complicated functions from the main interface. The remote will be made to look fashionable.

Figure 7.2: *Example decision points of a product design meeting.*

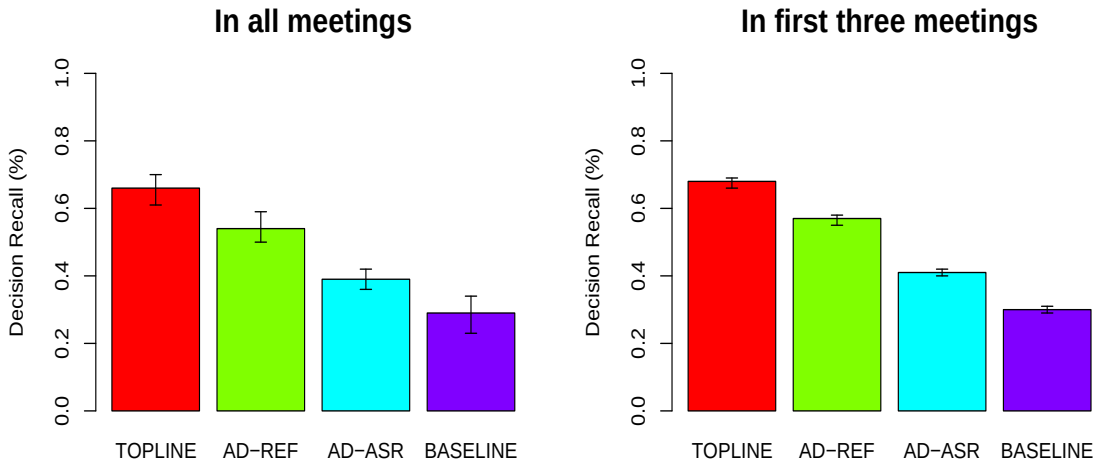

Figure 7.3: *Task effectiveness as the average ratio of the decisions that are correctly found by the subjects. These ratios are obtained from all meetings in the series (with a total number of 30 decision points) and from the first three meetings (with 24 decision points).*

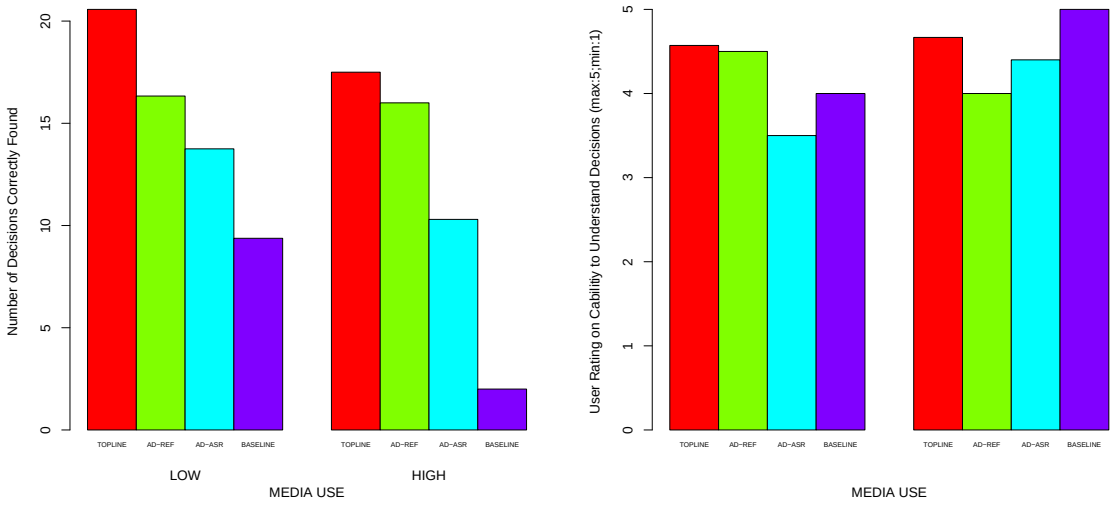

Figure 7.4: *Task effectiveness (number of decision hits) and perceived success (user ratings on understanding all decisions) as a function of media usage.*

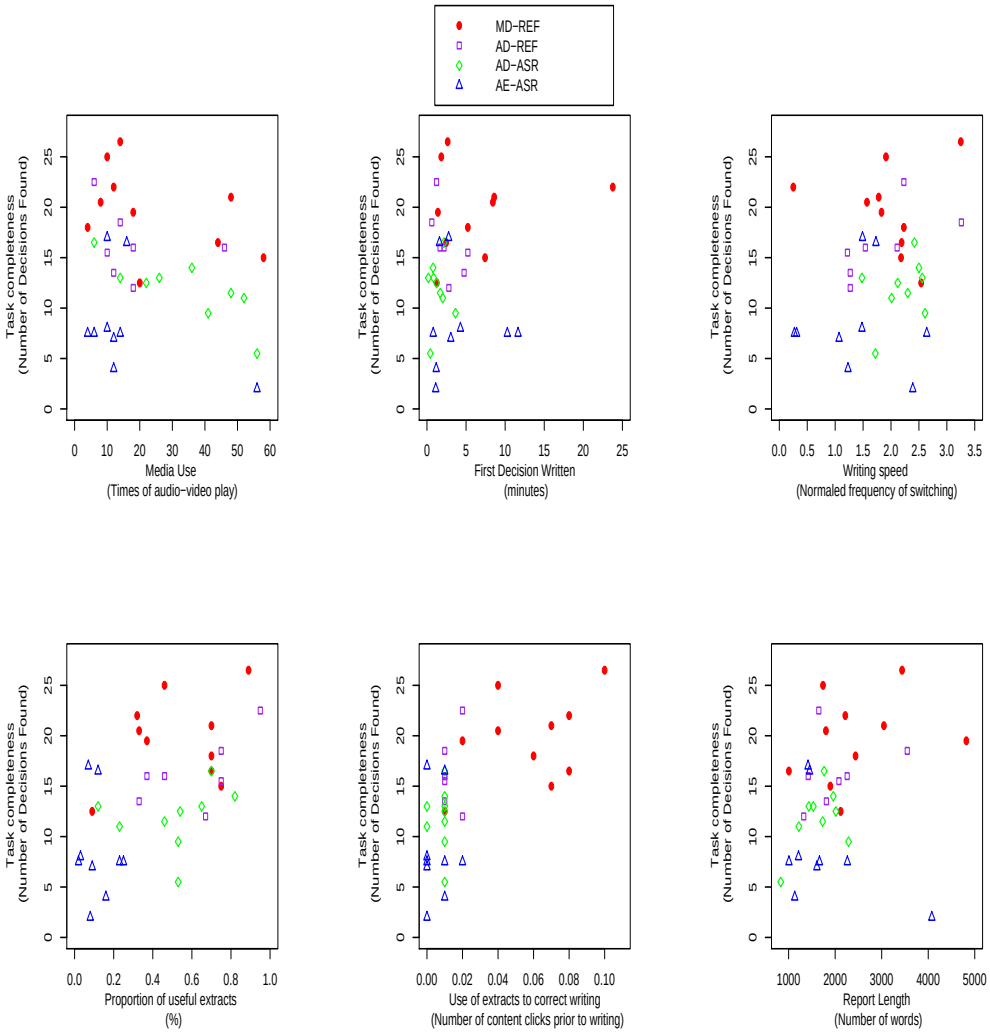

Figure 7.5: *Number of decisions found by the subjects from all the meetings as a function of log-based task effectiveness measures.*

Um so according to the brief um we're gonna be selling this remote control for twenty five Euro, And uh we don't want it to cost any more than uh twelve fifty Euros, so fifty percent of the selling price.
I mean do you think the fact that it's going to be sold internationally will have a bearing on how we design it at all?
Well right away I'm wondering if there's um th th uh, like with DVD players, if there are zones.
Well twenty five Euro, I mean that's um that's about like eighteen pounds or something,

Figure 7.6: *Example decision-focused extracts of a product design meeting.*

Chapter 8

Conclusion

The integration of lexical, multimodal, and multiparty cues enables the automatic derivation of decision-related discussions, which involves not only the detection of decision-related discussions, but also the determination of discussion segment boundaries. Studying these topics sheds light on the detection of other important types of high-level conversation phenomenon, such as problem-solving discussions and action item assignments. With the recent accumulation of multimedia recording archives, an automatic indexing tool that can map audio-video recordings into its conceptual organization for future use has become more and more important. The empirical studies and experiments reported in this thesis will serve as the first step towards the development of such an indexing tool.

Beyond the pragmatic benefits, this thesis contributes to the field of spoken language understanding from at least three aspects: (1) the understanding of the lexical, multimodal and multiparty cues of decision-related discussions and segment boundaries in meeting dialogues; (2) the development of computational means that enables the detection of these conversation phenomenon; and (3) the evaluation of the benefits brought by these computational means in real-life use scenarios. In the following sections, we review the contributions and limitations of each of these aspects and suggest possible future directions.

8.1 The Corpus Analysis of Decision-indicative and Discourse Segment-signalling Cues in Meeting Dialogues

The corpus analysis discussed in Chapter 4 constitutes a comprehensive empirical study of the decision-indicative cues in meeting dialogues. The database used in this analysis consists of 50 scenario-driven meetings (37,400 DAs) in which over 200 decisions are reached; 554 decision-related DAs were selected by the annotators to represent these decisions. The decision-indicative cues are uncovered from a variety of communication modalities, including the words spoken, gesture and head movements, pitch and energy level, rate of speech, pause, and use of subjective terms. The identified cues characterize the systematic differences of what the speakers say, how the participants express themselves, and how they interact with each other in decision-related discussions.

To our knowledge, this study is the first one that describes the decision-indicative cues in meeting dialogues. It is also the largest scale study that analyzes not only lexical cues, but also multimodal and multiparty cues hidden in the various communication modalities.

In addition to the decision-indicative cues, in Chapter 6 we also discussed the discourse segment-signalling cues in meeting dialogues, using the analysis framework established in the study of the decision-indicative cues.

8.1.1 The development of the lexical and multimodal feature selection framework for empirical analysis

In Chapter 2, we showed that the lack of proper computational models of natural multiparty discourse structures can be attributed to not only the variety of communication modalities used in a group context, but also the lack of a common framework that would integrate information from the multiple modalities. In Chapter 4, we first introduced a feature selection framework, which leverages statistical and information theoretic measures to quantify the discriminability of each of the lexical features (i.e., the occurrence of each N-gram in an analysis unit) to a certain conversation phenomena. The framework has been proven to be useful for picking out N-grams that are characteristic of the target conversation phenomena – that is, the decision-related discussions – for qualitative analysis. However, lexical information, which indicates what the speakers say, is complementary to multimodal and multiparty cues,, which reflects

how the speakers express themselves and how they interact with each other. To integrate information from the multiple knowledge sources, we extended the lexical feature selection framework, thus yielding a multimodal feature selection framework that selects discriminative features from heterogeneous knowledge sources.

The feature selection framework we proposed is independent of the discriminability measures. In this thesis, we experimented with both the statistical and information-theoretic measures of discriminability, with Chi-Squared statistics (X2) and Log Likelihood ratio (LL) in the first group and Dice Coefficient (DICE) and Point-wise Mutual Information (PMI) in the second group. The empirical analysis has shown that these measures capture similar subsets of features. The experiment of Section 5.8.1 has further demonstrated that all of these measures can successfully identify decision-characteristic features, except the PMI-based measure.

Figure 1.7 shows that how the decision detection task is positioned in the whole MDD system. The first step of multimodal source integration is the most time-consuming part. For example, the extraction of lexical features requires the existence of transcripts of spoken words (or the proxies of words, e.g., phonemes). At the time the experiments reported in this thesis were conducted, it would take at least 90-150 minutes (3-5 x real time) to obtain an ASR transcript from each of the 30-minute meetings using the AMI ASR system, which runs on models that were trained on all available data and adapted to the meeting domain and speakers. Compared to ASR, phoneme recognition is more time-efficient, saving the processing time to 0.5 x real time.

In addition, the extraction of prosodic features from the cross-talking recording of a 30-minute meeting will take another 10-20 minutes. This includes the time for processing fundamental freqency from speech, applying needed linearization to optimize the results, processing of energy-level information, and estimating the rate of speech.

Once the features are all readily extracted, the second step (i.e., decision detection) and the third step (i.e., discourse segmentation) can be done relatively fast, using the previously trained decision detection and discourse segmentation models. The processing of a 30-minute meeting for the two steps takes several minutes to complete.

While the processing time of meeting recordings throughout the three steps can be cut down to a great extent by replacing ASR with phoneme recognition, the training time of the models used in the process remains to be considerably high. Using the feature selection framework illustrated in Figure 5.4, the 1K+ features extracted from the multimodal inputs can be downsized to 6-18% of its original size. As the maximum entropy classification models can be optimized to run with linear complexity, the

feature selection framework can cut the training time by 82-94%.

8.1.2 The identification of key properties of decision-related discussions

In Chapter 4, we also provided a qualitative account of the key properties of decision-related discussions. In terms of the key lexical property, we analyzed the meeting transcripts and found that meeting participants mention more content words (e.g., *advanced chip*) and use fewer uncertain expressions (e.g., *I don't know*) in decision-related discussions. Additionally, they refer to the whole group (i.e., using the pronoun "we") more often than to themselves (i.e., "*I*") or to a specific group member (i.e., "*you*").

To identify other key multimodal properties, we analyzed the acoustics (e.g., pitch contour, energy level, and length of speaker pause) of the audio recordings and found decision characteristic cues in the way meeting participants express themselves. For example, meeting participants often stress their points with a higher than usual pitch and with a long pause, so as to capture the other participants' attention.

To identify key multiparty properties, we analyzed the annotated speaker intentions of each of the decision-related DAs and its immediate contexts, from where multiparty interaction patterns may emerge. The analyses show that decision-critical information is often expressed after participants have provided an evaluation of ideas (***assess***), or suggested an action related to the group or another individual (***suggest***). Moreover, meeting participants often reveal decision-critical information when they are providing information (***inform***) or expressing an action-related intention (***elicit-suggest***).

8.1.3 The identification of key properties of discourse segment boundaries

Because the detection of decision-related discussions also involves identifying the boundaries of the decision-related discussions. In doing so, we also conducted empirical analyses to determine the segment boundary-signalling properties in the following key aspects. We first analyzed the properties that signal the starting point or the end point of a discourse segment. The analyses indicate that meeting participants often use conventional expressions (e.g., *now*, *okay*, *lets*, *um*, *so*) to initiate a discussion; also, when discussing about decisions, they mention agenda items (e.g., *presentation*,

meeting) and topical words (e.g., *usability*, *costing*, *discussion*, *evaluation*) more than usual. While the use of the conventional expressions matches previous findings of spoken discourse segmentation, the use of topical and agenda-related words has not been studied in previous research.

Next, the analyses of meeting acoustics also yield segment boundary-signalling prosodic cues similar to previous findings. For example, a discourse segment is likely to start with higher pitched sounds (Brown et al., 1980; Ayers, 1994) and a lower rate of speech (Lehiste, 1980). In addition, participants often pause longer than usual to indicate the end of the current discussion (Brown et al., 1980; Passonneau and Litman, 1993).

The analyses of speaker intentions surrounding the context of discourse segment boundaries show additional boundary-signalling cues. For example, when attempting to initiate a new segment or to end the current discussion, meeting participants are often involved in the following states: providing information (***inform***), eliciting an assessment of what has been said so far (***elicit-assess***), starting speaking before they are ready (***stall***), or acting to smooth social functioning (***be-positive***).

Finally, as the AMI corpora comes with video recordings, we also analyzed the video recordings to understand head and hand/forearm movements among meeting participants, which were used to identify group action-based segments in previous research (McCowan et al., 2005; Al-Hames et al., 2005). The analyses show that meeting participants move around less than usual when a new discussion is brought up. In fact, head and hand/forearm movements and gestures were shown in previous research to be important modalities speakers use to convey meanings in face-to-face dialogues. For example, head nods are often used to indicate unsaid agreement or disagreement (Sacks et al., 1974; Cassell and Stone, 1999). Given that previously our analyses of the speaker behaviors simply considered the number of changed pixels over sliding windows, there is a great potential for spoken language researchers to further analyze the different aspects of speaker behaviors (captured in the video recordings) and to delve deeper into the problem of inferring group-based speaker intentions from multiparty interaction cues.

8.2 The Automatic Derivation of Decision-related Meeting Discussions

The corpus analysis provides empirical grounding for computing features and integrating them into a computational model. However, a qualitative account of the features provided by the empirical analysis is not sufficient for guiding the construction of the computational model that can recover decision-related DAs and hierarchical discourse segment boundaries. Here emerges the need of an algorithm that can automatically identify systematic patterns from a large set of features that might contain the identified decision-discriminative and segment boundary-signalling cues.

8.2.1 The integration of the lexical, multimodal, and multiparty information

In Chapter 5 and Chapter 6, we first trained a Maximum Entropy classifier to select and reweigh features in the model. The experimental results reported in Section 5.4 show that for adapting a supervised learning-based decision detector, lexical features would need to be combined with at least some of the multimodal or multiparty interaction features to achieve the best performance for identifying decision-related discussions from those that have been selected as important to understand overall meeting content, reducing the size of the extractive summaries to around 10%.

The leave-one-out analyses of the merits of feature classes suggest the following for the development of a well-performing automatic decision model: (1) Lexical features are essential to good performance on recall. (2) The additional non-lexical knowledge sources, e.g., the features that can characterize the meeting acoustics, dialogue act types, topics, and use of subjective language, are important to improving the precision of the model. When a model that issues fewer false positives is preferred, it is necessary to combine the lexical model with some of these knowledge sources, most effectively the dialogue act-based model, into a combined model. (3) When the trade-off between precision and recall is considered, combining the lexical features with more than one of the non-lexical knowledge sources has become necessary to achieve the best harmonic accuracy. Among the non-lexical knowledge sources, the prosodic features are the most important to be included, followed by the features indicating dialogue act types, topics, and use of subjective terms respectively.

Additionally, we also experimented with the task of detecting decision-related dis-

cussions from complete meeting recordings, the results in Section 5.5 further show that the lexical features are still the most important feature class in this task, albeit with an even greater positive impact.

Table 8.1 summarises how the different types of knowledge sources contribute to the automatic models.

	Decision Detection	Discourse Segmentation	
		TOP-LEVEL	ALL-LEVELS
Lexical			
Lexical			
Lexical			
Lexical			

Table 8.1: *Summary of the effect of multimodal knowledge sources on decision detection and discourse segmentation.*

8.2.2 Towards online processing

As ultimately the meeting dialogue understanding applications will be operated online or right after the end of a meeting, in the experiments reported in Section 5.5, Section 6.6.2 and Section 6.6.3 we also discussed two possible strategies to adapt the offline meeting decision detector: (1) Replacing the manually obtained knowledge sources with their automatically generated version, e.g., the ASR transcriptions generated by Hain et al. (2005) and the DA types predicted by the classifier developed in Dielmann and Renals (2007a), and (2) replacing word-based transcripts, of which the automatic generation process is costly and time-consuming, with sub-lexical, phoneme-based transcripts (or its enhanced version that further encodes speaker activity information), which can be generated in full automation.

The results show that, when moving online, the performance of the decision detection model will remain competitive with either strategy applied. For adapting an unsupervised, time series analysis-based segmenter that is incorporated in the decision detector, our novel phoneme (and speaker activity) enhanced approach can capture the multimodal and multiparty cues implicit in the group conversations and, in turn, improve the performance of the word-based segmenter originally used in the detector.

8.2.3 The advanced development of the feature selection framework

The main shortcoming of the simple feature selection framework discussed in Section 8.1.1 lies in its assumption that the discriminability of a feature to a target phenomena can be measured in isolation of other features. The formalization ignores the fact that conversation phenomenon are often characterized with patterns across multiple features, but not any single feature alone. Formalizing the feature patterns, e.g., with respect to feature intercorrelations or redundancy, and incorporating them into the models is therefore important to the development of the automatic meeting decision detector.

In Chapter 5, we compared the simple feature selection framework with that incorporating the consideration of feature patterns. Specifically, we evaluated the impacts of the frameworks on the efficiency and accuracy in detecting decision-related discussions. While the simple feature selection framework selects relevant feature subsets based on only their association (as measured in Chi-squared (X2) and information gain (IG)) with the decision-related discussions, the feature pattern-based framework leverages the measurements of feature intercorrelation and redundancy. In particular, the latter framework extends the former one and applies the following two algorithms:

- Correlation-based feature selection (CFS) (Hall, 1998): An advanced version of the X2 algorithm; Choose a subset of features that have the highest discriminability of decision-related DAs and the lowest intercorrelations among one another.
- Fast correlation-based filter (FCBF) (Yu and Hatzivassiloglou, 2003): An advanced version of the IG algorithm; Choose a subset of features that have the highest discriminability of decision-related DAs and remove those that are redundant.

In the experiments reported in Section 5.8, we examined the results yielded by the different feature selection criteria on the problem of detecting decision-related dialogue acts and decision-related discourse segments. The results were also compared using different feature combinations. For example, the LX models were trained with only the lexical features and the ALL models were trained with the combination of all available features, including lexical, prosodic, dialogue act type, topic, and subjective term features.

The comparison of the reduced feature sets shows that all criteria are effective in reducing feature space, thus increasing model efficiency. On average, the feature selection framework can reduce the LX model to around 11%-21% of its original size and the ALL model to around 14%-18%. When the feature pattern-based framework is applied, the model efficiency can be further improved, reducing 50%-66% of the feature subset selected by the feature selection methods previously used.

The comparison of the accuracy of models shows that although the reduced model is 5 to 15 times more efficient than the original model, it is only, at best, as effective as the original model. Further analysis suggests that applying feature selection is only beneficial to the precision of the decision detection models, but not the recall rate.

We also compared the accuracy of the models yielded by the two frameworks: single feature- and feature pattern-based one. Surprisingly, the results obtained with the pattern-based framework are not always better. Under the lexical model setting, the FCBF algorithm (which incorporates the minimum feature redundancy criterion in the feature selection process) improves on the precision of models for both the task of detecting decision DAs and that of detecting decision segments; the CFS algorithm (which incorporates the minimum feature intercorrelation criterion) achieves the best recall rates among the four, but only improves the precision on detecting decisions at the DA level. One possible explanation is that redundancies do exist among the lexical features, and filtering out the intercorrelated lexical features might have enforced the detection of decision-related DAs in some certain types of discourse segments.

Under the ALL model setting, the benefits of using FCBF on the precision for detecting decision-related DAs are cancelled out, suggesting the inclusion of multimodal and multiparty features is only detrimental to the selection of redundant features. Moreover, the benefits of using CFS on the recall rate of models are also cancelled out, and so are its benefits on the precision at the DA level. We posit that applying the additional feature pattern finding criteria on all of the heterogeneous features (including multimodal and multiparty features) might have caused the elimination of important lexical features that are representative of some decision-related discussions.

8.2.4 The advanced development of the knowledge source integration framework

Because we expect the different algorithms to capture features or patterns that are indicative of different types of decision-related discussions, we expect each of the

reduced models yielded by these algorithms to be capable of predicting only some certain types. Therefore, we also experimented with a knowledge source integration framework which aims to find synergy from the multiple reduced models. In particular, in the experiment of the integration framework reported in Section 5.9 we applied a simple ensemble construction algorithm, i.e. voting. The result shows that if we construct an ensemble model from these reduced models, we can improve the harmonic accuracy of the original model by an additional 10%, yielding 61% and 72% for the task of detecting decision-related DAs and segments respectively.

8.2.5 The task-based evaluation of decision-focused summary displays

In Chapter 5, we evaluated our meeting decision detection models intrinsically, using standard metrics. In Chapter 7, we described an extrinsic task-based evaluation of these models. Specifically, we evaluated the use of the decision-focused extractive summaries produced by the meeting decision detector for achieving the "decision debriefing" task.

This study verifies our experimental hypothesis: Existing meeting summarization systems, which aim to provide a general-purpose summary display, can be improved by refocusing the displayed summaries with regard to user's information need. For users who require a quick overview of meeting conclusions, the decision-focused summary display can improve not only the number of decisions recalled by the users, but also the quality of the meeting minutes they write.

Further analysis reveals that, when the displayed meeting summary contains ASR errors, the users rely more on the decision-focused display to summarize meeting decisions than the general-purpose summary. Also, as the summaries presented in the ASR-based display are error-prone, the audio-video recordings are used more frequently by the users during the process of accomplishing this task.

To lend support to the design of meeting browsers and the automatic machinery that generates query-focused speech summaries, we also investigated user behaviors during the use of speech summaries and analyzed whether the errors introduced by the automatically generated decision summaries (which consist about 30%-40% of the whole summaries) resulted in some differences in the user behaviors. The results show that although users prefer to leverage manual meeting summaries to find decisions, their perceived success towards the decision debriefing task and their ability to produce

quality minutes are not significantly affected. This confirms our intuition that users view the displayed summaries as pointers to the relevant discussions in the transcript.

8.3 Limitation and Future Direction

8.3.1 The liability of decision-based meeting summarization

Our decision DA classification approach views the decision summarization task as that of compiling an automatically selected set of DAs to represent meeting decisions (i.e., as an excerpt of the decisions). This approach has some inherent liabilities. First, the unconnected dialogue acts in the excerpt result in semantic gaps that require contextual information to bridge. Second, the anaphora and unexpected topic shifts between the extracted DAs also require contextual information to resolve. Last but not least, although it is our intuition that the decision-related DA extracts will assist users in finding and absorbing information in the meeting archives, this assumption has yet to be tested with human subjects.

To address the first liability and provide the needed contextual information, in Chapter 6 we trained computational models to determine segment boundaries, i.e., where a decision-related discussion is initiated and ended, in order to find relevant contexts for interpreting the identified decisions. The identified segments can also be used to automatically indicate the topic of the current decision-related discussion.

To address the second liability, in Chapter 7 we conducted an extrinsic evaluation of the decision audit task, in which the users were asked to summarize the decisions made in a series of AMI scenario-based meetings. In the following section, we review this evaluation, assessing the merit of displaying decision-related DA information (c.f. Figure 1.6) and that of displaying general-purpose extractive DA (c.f. Figure 7.1) to the users in the decision debriefing task that is common in daily organization lives.

8.3.2 Automatic detection of argument process and outcome

There are also directions with our decision DA classification approach we did not pursue in this thesis. First, as we relied on the annotators' judgement of what is a decision-related DA in the three-phase decision annotation procedure (as described in Chapter 3), the selected decision-related DAs by nature serve for various functional roles in a decision-making process.

Take the dialogue demonstrated in Figure 1.6 (see Figure 8.1 for the transcript of the decision discussion in the display) for example. The annotators marked dialogue act (9), (13), (16), and (44) as the decision-related DAs related to this decision: *"There will be no feature to help find the remote when it is misplaced"*. Among the four decision-related DAs, (9) describes the topic of what this decision is about; (13) and (16) describe the arguments that support the decision-making process; (44) indicates the level of agreement or disagreement for this decision.

In fact, the DAs belonging to each of the different function roles might be distinctive in their own characteristic features. However, in our approach, we trained one model to recognize DAs of the various functional roles. This approach might have degraded the performance of our model. As reviewed in Chapter 2, prior work suffered from the lack of understanding of the argumentative process in meetings; only a few schemes of argument structures have been attempted with automatic derivation. In particular, Rienks et al. (2005) attempted to automatically identify six classes of argument acts and nine classes of argument relations; Fernndez et al. (2008) followed prior work in action item identification Purver et al. (2006) to detect discourse units that serve four different roles, including issue, proposal, restatement, and agreement.

Although it is beyond the scope of this thesis, we expect similar studies of the major function roles that serve the progression of other types of discussions (e.g., problem solving, action item assignment) to be beneficial. For one, the lessons learned in the studies can shed light on the development of meeting dialogue understanding applications, in which the different stages of a meeting argumentative process, starting from the initial discussion of a problem, the possible solutions, the pro and con arguments towards each proposed solution, to the decision on how to resolve the problem. This future direction will require a large scale study on not only the key properties of the major discourse units contributing to the recognition of the different stages, but also the argument relations between these discourse units in the argument process.

Second, the current approach does not provide the capability of understanding the standing of each participant in the decision-making process. Advanced studies in how to distinguish private states (Quirk et al., 1985) in multiparty dialogues would be needed. Previous research in sentiment analysis studied the detection of private states by learning the subjective expressions that have been associated strongly with the private states in texts (Wiebe and Riloff, 2005; Pang and Lee, 2004; Yu and Hatzivassiloglou, 2003; Turney, 2002; Hatzivassiloglou and McKeown, 1997). However, the effectiveness of such techniques on distinguishing speaker states in spontaneous,

multiparty meeting speech has not been studied yet.

Third, because the target conversation phenomena (i.e. a decision-related DA) is usually a rare event, the success of our machine learning approach have been seriously impeded by the issue of an imbalanced class distribution. So far we tackled this issue by attempting with various sampling techniques (e.g., downsampling, upsampling); however, the results yielded from these attempts did not fare differently with the original condition on which no sampling was applied.

Last but not the least, our supervised learning approach requires a certain amount of annotated data to work, hence the availability of annotations would affect the applicability of this approach in other domains of natural dialogues. To increase the applicability of our approach, it would be interesting to explore semi-supervised approaches, in which the automatically annotated discourse units are combined with the manually annotated discourse units. The semi-supervised approaches would bootstrap training data from those unannotated one and, in turn, relax the requirement of the amount of data needed for training well-performing models. In addition, active learning and transfer learning techniques can also be applied to decrease the amount of annotations required. These techniques include the following: training models with labeled data available in other domains; training models in other domains, but only use the set of features that have been observed in the target domain; constructing ensemble models from the multiple models obtained from training in the different domains.

8.3.3 The Development of Meeting Dialogue Understanding Applications

With the algorithms of decision detection and discourse segmentation established, many different types of meeting dialogue understanding applications can be extended from this work. Among all, the most obvious ones are certainly tools that can automatically recover *where the decision-related dialogues are*.

In addition, we can also build tools that *provide relevant contexts* for interpreting the decisions made in the meetings. Our task-based evaluation shows that, when browsing meeting excerpts on a browser (c.f., Figure 1.3), the users often need to refer back to the relevant discussions in the transcription.[1] The key to providing pointers for semantic interpretation is to determine relevant neighboring discourse sequences.

[1]Our browser is designed to allow users to refer back to the relevant utterances in the meeting transcription by clicking on the sentence of interest in the meeting excerpt.

This solution calls for a better discourse segmentation mechanism, which we have developed in Chapter 6, to divide a complete meeting stream into a number of coherent segments.

Discourse segmentation is important since it can help organize the often-lengthy dialogues into a series of short segments, each with a coherent topic. The relatively shorter segments also lend support to the development of many meeting dialogue understanding applications. For example,

1. Information extraction (IE): for automatically extracting key lexical units that can describe the topic of each segment.

2. Summarization: for automatically generating a summary that consists of the gists of the segments.

3. Title generation: for automatically generating a title from the highlight of each segment.

4. Topic detection and tracking: for automatically clustering the topics of the segments into similar groups and organizing these segments into a hierarchical visual presentation that is easier for the users to browse (Allan et al., 1998).

With the multimodal feature selection and knowledge source integration frameworks established, many other downstream multimedia systems also benefit from the additional capability of inferring speakers' argumentative intents from various communication modalities. Take multimedia search for example. The automatically derived indices of speakers' argumentative intents can make up for the lack of proper indices, especially for what is expressed beyond words, in the multimedia archives. By combining the intent-based indices with the content-based ones, a better search facility can be developed. This is because the combined indices offer an integrated view of what is truly going on, not just what has been said in the conversations.

Take human-computer dialogue systems as another example. Previous research largely focuses on developing systems that help convey information, e.g., the flight and hotel booking information, direction to a certain location, descriptions of properties for rent or sale. However, these dialogue systems do not have the capability of acknowledging what the users intend to argue. It is therefore an interesting future direction to study how to better incorporate an argumentative intent analysis component into dialogue systems and how to use the new component to guide the generation of proper responses.

========== Beginning-of-Discussion ==========
(9) A: But um the feature that we considered for it not getting lost.
(10) B: Right. Well.
(11) B: We're talking about that a little bit.
(12) B: When we got that email and we think that each of these are so distinctive,
(13) B: that it it's not just like another piece of technology around your house.
(14) B: It's gonna be somewhere that it can be seen.
(15) A: Mm-hmm.
(16) B: So we're we're not thinking that it's gonna be as critical to have the loss.
(17) D: But if it's like under covers or like in a couch you still can't see it.
(18) A: Its really
(19) A: Would it be very difficult to um just have an external device
(20) A: like i dunno you tape to your to your TV
(21) A: Um that when you press it
(22) A: you ha
(23) A: a little light beep goes off
(24) A: Do you think that would be conceptually possible
(25) C: I think it would be difficult technologically
(26) B: I think
(27) A: Mm-hmm
(28) C: Because if your if your remotes lost its probably under the settee
(29) C: And in that case you can't send an infrared sing signal to it to find it
(30) A: Mm
(31) C: So its
(32) C: I'm not quite sure how it would work
(33) A: That's true mm kay
(34) B: Yeah.
(35) C: And then i wonder if it's if it's more just a gimmick then anything else
(36) C: Uh i mean
(37) C: Ho how many times do you really seriously lose your remote control
(38) C: And would would a device like that actually help you to find it
(39) B: There might be something that you can do in the circuit board and the chip to make it make a noise or something
(40) B: But it would take a lot more development than we have this afternoon
(41) A: Mm-hmm
(42) A: Okay, that's a fair evaluation.
(43) A: Getting lost.
(44) A: Um we so we do we've decided not to worry about that for now.
(45) A: Okay.
(46) A: Cause well the designs are very bright.
(47) A: So you're right. They're gonna stick out.
========== End-of-Discussion ==========
(48) A: but um
(49) B: so d do people have a preference as far as feel and functionality um
(50) D: i feel like this is simil or its sort of what already exists

Figure 8.1: *Example decision-making discussion*

Appendix A

Appendix A: 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) ?

Appendix B: Task Instructions for Decision-fcused Extract Subjects (in Condition Topline, AD-REF, and AD-ASR)

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. The four participants include: the 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. In addition, for each meeting you are presented with a list of sentences that are considered to be important to interpret the DECISIONS that were made in that meeting. These sentences were selected from the transcript. Therefore, clicking on one of those sentences takes you to that sentence's position in the meeting transcript and where the meeting participants are talking about it in the audio-visual recordings. It is possible that not all of the decisions are captured in the list of sentences.

In this study, we are interested in the group's decision-making ability, and therefore ask you to evaluate and summarize this particular aspect of their discussion. Please imagine the following scenario: Your manager, who has missed these meetings, has asked you to summarize the decisions made in the meetings for her. The group discussed a range of issues regarding the remote control they are designing.

Please include the following information in your report: A list of the decisions made in each meeting. (You can do this as bulleted lists). Do keep in mind that your report will be part of the project documentation. It is therefore essential that you keep your report short and of high quality. Do not use any abbreviations, as your manager may not have sufficient background information to interpret them.

Please write your summary in the browser tab labelled "Typing tab."

You have a total of 45 minutes for this task. As on average each meeting lasts around 35 minutes and there are four meetings, please do leave yourself enough time to complete the written summary. I will give you two warnings when there are 20 and 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.

Appendix C: Task Instructions for General-purpose Extract Subjects (in Condition Baseline)

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. The four participants include: the 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. In addition, for each meeting you are presented with a list of sentences that are considered to be important to interpret what has happened in that meeting. These sentences were extracted from the transcript. Therefore, clicking on one of those sentences takes you to that sentence's position in the meeting transcript and where the meeting participants are talking about it in the audio-visual recordings.

In this study, we are interested in the group's decision-making ability, and therefore ask you to evaluate and summarize this particular aspect of their discussion. Please imagine the following scenario: Your manager, who has missed these meetings, has asked you to summarize the decisions made in the meetings for her. The group discussed a range of issues regarding the remote control they are designing.

Please include the following information in your report: A list of the decisions made in each meeting. (You can do this as bulleted lists). Do keep in mind that your report will be part of the project documentation. It is therefore essential that you keep your report short and of high quality. Do not use any abbreviations, as your manager may not have sufficient background information to interpret them.

Please write your summary in the browser tab labelled "Typing tab."

You have a total of 45 minutes for this task. As on average each meeting lasts around 35 minutes and there are four meetings, please do leave yourself enough time to complete the written summary. I will give you two warnings when there are 20 and 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.

Appendix D: Post-questionnaire

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. I understood the decisions made in the meeting discussions. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

7. The task required a great deal of effort. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

8. I had to work under pressure. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

9. I had the tools necessary to complete the task efficiently. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

10. I would have liked additional information about the meetings. (disagree strongly 1-2-3-4-5 agree strongly)

answer=

11. 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 list of important sentences about decisions from each meeting?

12. What information would you have liked that you didn't have?

Appendix E: Gold Standard Decision-focused Abstracts in the ES2008 Series

ES2008A:

- One remote will have main functions (on, off, channel changing, and volume)(basic, everyday functions)
- Another will have special (complicated) functions
- Large buttons
- Easy to use
- The remote will be sold for about 25 Euro, The profit aim will be 15 million Euro, The production cost will be a maximum of 12.50 Euro
- It will target be an international market

ES2008B:

- Rather than define a specific target group by age, the team will define objectives such as fashion and simplicity instead
- The team will not work with teletext
- The remote will be used only with televisions
- The corporate image must be recognizable on the remote
- The team's design will focus on simplicity and fashion
- The remote will contain a (locator) function to aid in its recovery when lost
- The remote will have (large) buttons for essential functions
- The remote will have a possibility to get extra functions
- The buttons for extra, infrequently used functions will be hidden (less visible) in the design in some manner
- The remote will have a rechargeable battery (energy source) and a charging station

ES2008C:

- The target group comprises of individuals who can afford the product
- The remote will use a kinetic battery
- The remote will have a latex case
- The remote will be made in fruity colors
- The remote will have some sort of a (fashionable) curved case
- The remote will have pushbuttons
- The remote will have a power button, volume buttons, channel preset buttons, and a menu button
- The remote will have a simple chip

ES2008D:

- The remote will resemble the potato prototype
- There will be no feature to help find the remote when it is misplaced; instead the remote will be in a bright color to address this issue
- The corporate logo will be on the remote; one of the color concepts for the remote will contain the corporate colors
- The remote will have six buttons;
- The buttons will all be one color
- The case will be single curve, made of rubber, and have a special color

Appendix F: Gold Standard Decision-Focused Extracts in the ES2008 Series

ES2008A

START	END	EXTRACTED TEXT
455.49	463.13	we'd like to sell it for about twenty five Euro with the profit aim of um fifteen million Euro um from our sales
463.13	469.42	and because this is such this is for television it's a we have a market range of Internet , like it's an international market range ,
471.89	482.36	Um in order to make a profit of this magnitude , we need to um be able to produce each one at a maximum of twelve fifty Euro .
721.09	728.23	but that just has your major buttons for that work for everything , you know volume control , on , off ,
752.48	760.24	and if um if you'd save the more complicated functions maybe for separate remotes that you wouldn't need to use every day .
761.49	768.66	so maybe have like one remote that has the main functions on , off , channel changing , volume , and another rote remote with all the special things .

ES2008B

1281.27	1283.31	but we're not gonna work with teletext
1365.34	1372.54	like your question earlier um whether this is going to be t for television , video , or etcetera . Just for television . That's what we're focused on .
1379.89	1385.99	Um and finally there's more marketing , I think , um , our corporate image has to be recognisable .
1386.28	1391.16	while we're gonna make it look pretty we need to use our colour and our slogan i in the new design .
1659.80	1668.32	So if you wanna do something complicated like programme your television or re-tune it , then you you open up this little hatch or or slide the screen down
1668.32	1670.04	and there's all the all the special buttons .
1705.74	1709.8	what are we emphasising ? I what in this project ?
1711.32	1712.87	I think simplicity , fashion .
1875.80	1878.67	maybe we don't have to defi define the target group by the demographic of age ,
1878.67	1880.78	maybe we can define it by like the demographic of
1882.29	1885.18	like h t how much money they have to spend or something like that ,
1893.05	1901.76	So maybe it's more useful to d d to define objectives like fashion and simplicity than to find specific target group as far as age is
1911.32	1914.93	do we want some kind of thing to find it if it's lost ?
1916.53	1918.25	Like a button on a T_V_ you can press
1941.83	1944.04	It would be relevant to like the overall goal I think ,
1951.13	1953.55	Um so we want something to keep it from getting lost .
1956.28	1965.24	And we want um we want large buttons for the essential things .
1972.47	1981.95	We want a possibility to um to get um a possibility to get the extra functions .
1985.90	1988.36	Which are kind of hidden away in some way
2055.75	2061.98	But you could maybe have it in a little charging station like a mobile phone , or like a little cradle for your iPod .
2065.83	2068.32	So you don't ha you got like a rechargeable battery .

ES2008C

1614.21 1617.09	what do people think about this kinetic battery idea ?
1616.90 1619.22	I think it's awesome . I think it's really cool .
1622.79 1627.59	it would t totally take care of our problem of not wanting to change batteries .
1828.53 1833.62	Uh I'm kinda liking the idea of latex , if if spongy is the in thing .
1848.22 1859.42	what I've seen , just not related to this , but of latex cases before , is that there's uh like a hard plastic inside , and it's just covered with the latex .
1867.43 1870.32	I don't think we need to worry about protecting the circuit board ,
1880.19 1884.8	Um and probably in colours , maybe fruity , vegetable colours .
1896.3 1897.72	and we want a curved case ,
1898.49 1899.75	Or a double-curved ?
1902.21 1904.48	I'm thinking curved of some sort .
1926.7 1928.88	'cause then we can have a a simple chip
1934.13 1939.4	So in terms of uh in terms of uh economics it's probably better to have pushbuttons .
1985.90 1992.45	but if it's gonna be in a latex type thing and that's gonna look cool , then that's probably gonna have a bigger impact than the scroll wheel .
2021.31 2023.08	what are what are our buttons gonna be ?
2023.08 2024.17	On off
2024.84 2030.92	uh volume , favourite channels , uh and menu .
2079.47 2081.49	Let's go with a simple chip ?

ES2008D

339.26	346.64	in terms of making decisions , what we'd need to do is first of all decide on a form uh which of the three different shapes we want ,
384.97	388.16	something still a little bright to make it hard to lose , but
395.32	399.77	but um the f the um feature that we considered for it not getting lost .
404.33	411.05	and we think that each of these are so distinctive , that it it's not just like another piece of technology around your house .
411.05	414.15	It's gonna be somewhere that it can be seen .
479.65	482.61	Um we so we do we've decided not to worry about that for now .
610.08	612.93	is that where people are leaning then , the potato ?
613.52	614.87	I think I'm leaning towards the potato .
614.86	618.33	that's really gotten the simplicity of the buttons down , that one .
675.13	683.24	which one are we sort of roughly looking at to address whether or not it meets our s um necessities ,
684.81	686.78	The potato ? Are we leaning towards the potato ?
934.95	936.37	is it gonna be yellow ?
937.56	938.28	It it might be ,
938.28	939.34	'cause that's our corporate colour ,
943.22	944.37	We might wanna keep it yellow .
948.83	953.80	but if we had all the buttons in black , and a design in and the outside in yellow , that'd be our corporate one
962.34	965.77	Um and can we have like an R_R_ inscribed on the bottom or something ?
1256.81	1260.44	we didn't we didn't address the fact that it does need to b have a corporate logo ,
1260.44	1267.10	so let's let's make sure we keep that in mind that we ha that one of our colours concepts is corporate and has an R_R_ on it .
1398.26	1399.81	we're doing push buttons .
1400.61	1402.47	We have six .
1476.82	1478.75	let's have our buttons all be one colour .
1488.24	1489.61	are we sure this is double-curved ?
1489.61	1491.41	Maybe it's single-curved ,
1498.92	1501.07	It's single curved .
1508.16	1510.75	but we have a simple chip , single curve ,
1512.22	1513.62	case material is rubber
1513.62	1516.70	and it's a special colour ,
1518.51	1521.63	Six buttons we have to have six buttons .

Appendix F: Gold Standard Decision-Focused Extracts in the ES2008 Series

DECISION_DISCRIMINATIVE_WORDS_SELECTED_BY_LOGLIKELIHOOD

WORD	LL-RANK	LL-SCORE	FREQ_IN_DEC	FREQ	DEC_FREQ	CONTENT_WORD?
yeah	1	247.79	21	6068	6017	0
mm	2	83.66	4	1806	6017	0
okay	3	73.23	18	2625	6017	0
we	4	52.68	212	4090	6017	0
buttons	5	39.92	56	704	6017	1
case	6	30.88	20	145	6017	1
oh	7	27.4	3	732	6017	0
advanced	8	26.98	13	70	6017	1
curved	9	26.97	15	94	6017	1
display	10	26.88	14	82	6017	1
rubber	11	26.63	21	180	6017	1
flip	12	23.93	9	36	6017	1
slogan	13	23.2	6	14	6017	1
selling	14	22.95	10	48	6017	1
station	15	22.82	12	71	6017	1
cost	16	22.32	18	157	6017	1
teletext	17	21.31	15	117	6017	1
decided	18	20.7	10	54	6017	0
euro	19	20.63	11	66	6017	1
you	20	20.17	105	5104	6017	0
weve	21	19.79	21	223	6017	0
chip	22	19.45	20	208	6017	1
docking	23	19.14	7	27	6017	1
volume	24	17.96	18	184	6017	1
euros	25	17.63	11	77	6017	1
vegetable	26	17.63	7	30	6017	1
and	27	16.32	206	5050	6017	0
production	28	14.93	9	61	6017	1
be	29	14.65	96	2084	6017	0
concentrate	30	14.31	3	5	6017	1
battery	31	14.12	14	142	6017	1
colours	32	14.01	16	178	6017	1
curving	33	13.94	2	2	6017	1
shuts	33	13.94	2	2	6017	1
motion	33	13.94	2	2	6017	1
folded	33	13.94	2	2	6017	1
colour	34	13.93	17	197	6017	1
recognition	35	13.78	18	217	6017	1
im	36	13.73	5	592	6017	0
l_c_d_	37	13.38	19	240	6017	1
corporate	38	13.29	6	30	6017	1
image	39	13.29	5	20	6017	1
digits	40	13.1	4	12	6017	1
countries	40	13.1	4	12	6017	1
were	41	12.84	44	804	6017	0
international	42	12.21	6	33	6017	1
maybe	43	12.17	41	745	6017	0
microphone	44	11.99	7	46	6017	1

DECISION_DISCRIMINATIVE_WORDS_SELECTED_BY_LOGLIKELIHOOD

fifty	45	11.56	10	92	6017	1
world	46	11.47	5	24	6017	1
bit	47	11.36	1	285	6017	0
should	48	11.25	34	596	6017	0
thats	49	10.72	20	1242	6017	0
recognisable	50	10.64	3	8	6017	1
changeable	50	10.64	3	8	6017	1
my	51	10.45	3	403	6017	0
single	52	10.43	8	67	6017	1
think	53	10.28	80	1799	6017	0
voice	54	10.26	14	173	6017	1
india	55	10.18	2	3	6017	1
soccer	55	10.18	2	3	6017	1
maximal	55	10.18	2	3	6017	0
ruled	55	10.18	2	3	6017	0
imprint	55	10.18	2	3	6017	1
indicator	55	10.18	2	3	6017	0
i	56	10.18	123	5243	6017	0
here	57	9.75	3	388	6017	0
but	58	9.72	36	1880	6017	0
menu	59	9.68	11	122	6017	1
print	60	9.61	6	42	6017	1
yes	61	9.61	2	323	6017	0
plastic	62	9.58	10	105	6017	1
d_v_d_	63	9.38	5	30	6017	1
cause	64	9.37	2	318	6017	0
away	65	9.35	8	73	6017	0
could	66	9.16	39	760	6017	0
incorporate	67	9.08	5	31	6017	0
project	68	9.01	1	242	6017	0
ed	69	8.86	4	20	6017	0
add	70	8.85	8	76	6017	0
covers	71	8.8	5	32	6017	1
right	72	8.77	12	820	6017	0
down	73	8.7	16	230	6017	0
price	74	8.69	9	94	6017	1
better	75	8.66	10	112	6017	0
illuminate	76	8.52	2	4	6017	0
sensitive	76	8.52	2	4	6017	1
cup	76	8.52	2	4	6017	1
mistake	76	8.52	2	4	6017	0
maximum	76	8.52	2	4	6017	0
speech	77	8.3	12	153	6017	1
include	78	8.07	6	49	6017	0
speaker	78	8.07	6	49	6017	1
power	79	8.06	10	117	6017	1
make	80	8.05	29	538	6017	0
next	81	8.04	1	224	6017	0
there	82	8.03	8	610	6017	0

DECISION_DISCRIMINATIVE_WORDS_SELECTED_BY_LOGLIKELIHOOD

yep	83	7.99	1	223	6017	0
under	84	7.87	6	50	6017	0
cells	85	7.81	4	23	6017	1

Bibliography

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

Allan, J. (2002). *Introduction to topic detection and tracking*. Kluwer Academic Publishers, Norwell, MA, USA.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1):125–127.

Almuallim, H. and Dietterich, T. G. (1991). Learning with many irrelevant features. In *Proceedings of the Ninth National Conference on Artificial Intelligence*.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialogue. In *IEEE Proceedings of ICSLP 2002*.

Ang, J., Liu, Y., and Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2005)*, Philadelphia PA USA.

Argyle, M., Ingham, R., Alkema, F., and McCallin, M. (1973). The different functions of gaze. *Semiotica*, 7:19–32.

Arons, B. (1994). Pitch-based emphasis detection for segmenting speech recording. In *Proceedings of International Conference on Spoken Language Processing*, volume 4, pages 1931–1934.

Aue, A. and Gamon, M. (2005). Customizing sentiment classifiers to new domains: a case study. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*.

Ayers, G. M. (1994). Discourse functions of pitch range in spontaneous and read speech. In Venditti, J. J., editor, *OSU Working Papers in Linguistics*, volume 44, pages 1–49.

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

Banerjee, S. and Rudnicky, A. (2007). Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *Proceedings of IUI 2007*.

Banerjee, S. and Rudnicky, A. I. (2006). You are what you say: Using meeting participants' speech to detect their roles and expertise. In *Proceedings of the Workshop of HLT-NAACL 2006: Analyzing Conversations in Text and Speech*. ACM Press.

Barzilay, R. (2003). *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. PhD thesis, Columbia University.

Bates, M. J. (1990). The berry-picking search: user interface design. In Thimbleby, H., editor, *User Interface design*. Addison-Wesley.

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Noth, E. (2003). How to find trouble in communication. *Speech Communication*.

Baucells, M. and Sarin, R. K. (2003). Group decisions with multiple criteria. *Management Science*, 49(8):1105–1118.

Beach, L. and Mitchell, T. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, 3:439–449.

Beattie, G. and Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speec. *Journal of Language and Social Psychology,*, 18(4).

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

Berger, A., Pietra, S. D., and Pietra, V. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):3971.

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

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

Bodily, S. (1979). A delegation process for combining individual utility functions. *Management Science*, 25(10).

Bolt, R. A. (1980). Put-that-there: voice and gesture at the graphics interface. *Computer Graphics*, 14(3):262270.

Boulis, C. and Ostendorf, M. (2005). A quantitative analysis of lexical differences between genders in telephone conversation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. ACM Press.

Brown, G., Currie, K. L., and Kenworthe, J. (1980). *Questions of Intonation*. University Park Press.

Brown, M. G., Foote, G. J. F., Jones, K. S., and Jones, S. J. Y. (1995). Automatic content-based retrieval of broadcast news. In *Proceedings of ACM Multimedia 1995*, pages 35–43.

Brown, M. G., Foote, J. T., Jones, G. J. F., Jones, K. S., and Young, S. J. (1996). Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings of ACM Multimedia 1996*.

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

Busemeyer, J. R. and Diederich, A. (2002). Survey of decision field theory. *Mathematical Social Sciences*, 43:345–370.

CALO (2006). Cognitive agent that learns and organizes. *http://www.ai.sri.com/project/CALO*.

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

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

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

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., and Rich, C. (2001). Non-verbal cues for discourse structure. In *Annual Conference of the Association for Computational Linguistics (ACL 2001)*.

Cassell, J. and Stone, M. (1999). Living hand to mouth: Psychological theories about speech and gesture in interactive dialogue systems. In Susan Brennan, A. G. and Traum, D., editors, *AAAI 1999 Fall Symposium on Psychological Models of Communication in Collaborative Systems*.

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

Chaudhri, V. K., Cheyer, A., Guili, R., Jarrold, B., Myers, K. L., and Niekrasz., J. (2006). A case study in engineering a knowledge base for a personal assistant. In *Semantic Desktop and Social Semantic Collaboration Workshop at the International Semantic Web Conference*.

Chiclana, F., Herrera, F., and Herrera-Viedma, E. (2002). A note on the internal consistency of various preference representations. *Fuzzy Sets Syst.*, 131(1):75–78.

Chino, T. and Tsuboi, H. (1996). A new discourse model for spontaneous spoken dialogue. In *Proceedings of ICSLP 1996*.

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

Christensen, H., Gotoh, Y., Kolluru, B., and Renals, S. (2003). Are extractive text summarisation techniques portable to broadcast news? In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003)*.

Christensen, H., Kolluru, B., Gotoh, Y., and Renals, S. (2005). Maximum entropy segmentation of broadcast news. In *Proceedings of ICASSP 2005*, Philadelphia USA.

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

Clark, H. (1996). *Using Language*. Cambridge University Press.

Clark, H. and Schober, M. (1992). *Arenas of Language Use*, chapter Understanding by addressees and overhearers., pages 176–197. Chicago: University of Chicago.

Clemen, R. (1997). *Making Hard Decisions: An Introduction to Decision Analysis*. Wadsworth Publishing Company.

Cohen, M. S. (1993). *The naturalistic basis of decision biases*, chapter 3, pages 51–99. Norwood, NJ: Ablex.

Cohen, S. (2003). A computerized scale for monitoring levels of agreement during a conversation. In *Penn Working Papers in Linguistics Vol 8:1*, Philadelphia USA.

comScore Inc. (2007). comscore video metrix report.

Cooper, W., Paccia, J., and LaPointe, S. (1978). Hierarchical coding in speech timing. *Cognitive Psychology*, 10:154–177.

Core, M. G. and Allen, J. F. (1997). Coding dialogues with the damsl annotation scheme. In Traum, D., editor, *Working Notes: AAAI 1997 Fall Symposium on Communicative Action in Humans and Machines*, page 2835. American Association for Artificial Intelligence.

Cremers, A. H., Hilhorst, B., and Vermeeren, A. P. (2005). What was discussed by whom, how, when and where? personalized browsing of annotated multimedia meeting recordings. In *Proceedings of HCI*, pages 1–10.

Cutler, A., Dahan, D., and van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40:141–201.

Darroch, J. and Ratcli, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43:1470–1480.

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 74–81. Morgan Kaufmann Publishers Inc.

Davidson, D., Suppes, P., and Siegel, S. (1957). *Decision making: An experimental approach*. Stanford, CA: Stanford University Press.

de Finetti, B. (1964). *Studies in Subjective Probability*, chapter Foresight: Its logical laws, its subjective sources. NY: Wiley.

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

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193222.

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

Dielmann, A. and Renals, S. (2007a). Automatic meeting segmentation using dynamic bayesian networks. *IEEE Transactions on Multimedia*, 9(1):25–36.

Dielmann, A. and Renals, S. (2007b). DBN-based joint dialogue act recognition of multiparty meetings. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*.

Donaldson, G. and Lorsch, J. W. (1983). *Decision making at the top*. New York: Basic Books.

Dougherty, D., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretisation of continuous features. In *Proceedings of the Twelvth International Conference of Machine Learning (ICML 1995)*.

Dumais, S., Cutrell, E., and Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 2001)*, pages 277–284.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Prsonality and Social Psychology*, 23(2):283–292.

Duncan, S. and Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10:234–247.

Edmundson, H. P. (1968). New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Erol, B., Lee, D.-S., and Hull, J. (2003). Multimodal summarization of meeting recordings. In *Proceedings of 2003 International Conference on Multimedia and Expo (ICME 2003)*.

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

Fernndez, R., Frampton, M., Ehlen, P., Purver, M., and Peters, S. (2008). Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

Ferreira, F. and Bailey, K. G. D. (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, 8(5):231–237.

Forbes-Riley, K. and Litman, D. J. (2006). Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In *Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2006)*.

Fowler, C. & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26:489–504.

Furnas, G. W. (1986). Generalized fisheye views. *SIGCHI Bull.*, 17(4):16–23.

Furui, S., Hirohata, M., Shinnaka, Y., and Iwano, K. (2005). Sentence extraction-based automatic speech summarization and evaluation techniques. In *Proceedings of Symposium on Large-Scale Knowledge Resources*.

Fussell, S. R. and Kreuz, R. J. (1999). *Social and Cognitive Approaches to Interpersonal Communication: Introduction and Overview*, chapter 1. Lawrence Erlbaum Associates.

Galley, M., McKeown, J., Hirschberg, J., and Shriberg, E. (2004). Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the ACL (ACL 2004)*.

Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the ACL (ACL 2003)*.

Garofolo, J. S., Auzanne, C. G. P., and Voorhees., E. M. (2000). The trec spoken document retrieval track: A success sotry. In *Proceedings of RIAO 2000*.

Garofolo, J. S., Laprun, C. D., Michel, M., Stanford, V., and Tabassi, E. (2004). The NIST meeting room pilot corpus. In *Proceedings of LREC*.

Gatica-Perez, D., McCowan, I., Zhang, D., and Bengio, S. (2005). Detecting group interest level in meetings. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)*.

Gavalda, M., Zechner, K., and Aist, G. (1997). High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Proceedings of the Fifth ANLP Conference*, pages 12–15.

Georgescul, M., Clark, A., and Armstrong, S. (2006). An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In *The 7th SIGdial Workshop on Discourse and Dialogue*, pages 144–152.

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

Gordon, A. S. and Ganesan, K. (2005). Automated story extraction from conversational speech. In *Proceedings of the Third International Conference on Knowledge Capture (K-CAP)*.

Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., and Kajarekar, S. (2006). Combining prosodic, lexical and cepstral systems for deceptive speech detection. In *Proceedings of IEEE ICASSP 2006*.

Grice, H. P. (1969). Utterer's meaning and intentions. *Philosophical Review*.

Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics III: Speech Acts*, pages 41–58. New York: Academic Press.

Grimes, J. (1975). *The thread of discourse*. The Hague.

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

Grosz, B. and Kraus, S. (1993). Collaborative plans for group activities. In *Proceedings of IJCAI 1993*, pages 367–373, Chambery, France.

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

Gruenstein, A., Niekrasz, J., and Purver, M. (2005). Meeting structure annotation: Data and tools. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*.

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

Hall, M. (1998). *Correlation-based Feature Selection for Machine Learning*. PhD thesis, Waikato University, Department of Computer Science.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. London: Longman.

Harris, Z. S. (1963). *Discourse Analysis: Reprint*. The Hague, Mouton.

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

Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35st Annual Meeting of the Association for Computational Linguistics (ACL 1997)*.

Hauptmann, A. G. (2006). Trecvid: the utility of a content-based video retrieval evaluation. In *IS&T/SPIE Symposium on Electronic Imaging (EI)*, pages p. 6061–08.

Hauptmann, G. and Witbrock, M. J. (1997). Informedia: News-on-demand multimedia information acquisition and retrieval. In *Intelligent Multimedia Information Retrieval*, pages 213–239. AAAI Press.

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

Henderson, J. (2004). *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. Psychology Press.

Herrera, F., Herrera-Viedma, E., and Verdegay, J. (1995). a sequential selection process in group decision making with a linguistic assessment approach. *International Journal of Information Sciences*, 80:1–17.

Hillard, D., Ostendorf, M., and Shriberg, E. (2003). Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT-NAACL 2003*.

Hirschberg, J., Choi, J., Nakatani, C., and Whittaker, S. (1998). " i just played that a minute ago!:" designing user interfaces for audio navigation. In *Workshop On Content Visualization And Intermedia Representations*.

Hirschberg, J. and Litman, D. (1987). Now let's talk about now: identifying cue phrases intonationally. In *Proceedings of ACL 1987*, pages 163–171.

Hirschberg, J. and Nakatani, C. H. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of ACL 1996*.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, 3:67V90.

Hobbs, J. R. (1985). On the coherence and structure of discourse. Technical Report Report No. CSLI-85-37, Center for the Study of Language and Information Stanford University.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*.

Howard, R. and Matheson, J. (1984). *Readings on the Principles and Applications of Decision Analysis*, volume 2, chapter Influence diagrams. Menlo Park CA: Strategic Decisions Group.

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

Hsueh, P. and Moore, J. D. (2007a). Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the 45th Annual Meeting of the ACL (ACL 2007)*. Association for Computational Linguistics.

Hsueh, P. and Moore, J. D. (2007b). What decisions have you made?: Automatic decision detection in meeting conversations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 25–32, Rochester, New York.

Hurst, M. and Nigam, K. (2004). Retrieving topical sentiments for online document collections. In *Proceedings of Document Recognition and Retrieval (DCR 2004)*.

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

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203.

Jaimes, A., Omura, K., Nagamine, T., and Hirata, K. (2004). Memory cues for meeting video retrieval. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 74–85.

Janin et al., A. (2003). The ICSI meeting corpus. In *Proceedings of ICASSP 2003*.

Jensen, F. (2001). *Bayesian Networks and Decision Graph*. Springer.

Ji, G. and Bilmes, J. (2005). Dialog act tagging using graphical models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP 2005)*.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, page 137142.

Johnston, M. and Bangalore, S. (2000). Finite-state multimodal parsing and understanding. In *Proceedings of the International Conference on Computational Linguistics (COLING 2000)*.

Jones, G., Foote, J., Jones, K. S., and Young, S. (1997). *The video mail retrieval project: experiences in retrieving spoken documents*, chapter 10, pages 191–214. MIT Press, Cambridge, MA, USA.

Jovanovic, N., op den Akker, R., and Nijholt, A. (2006). Addressee identification in face-to-face meetings. In *the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Jurafsky, D., Bates, B., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and Ess-Dykema, C. V. (1997). Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of IEEE Workshop on Speech Recognition and Understanding*.

Kacprzyk, J. (1986). Group decision making with a fuzzy linguistic majority. *Fuzzy Sets Syst.*, 18(2):105–118.

Kan, M. (2003). *Automatic text summarization as applied to information retrieval: Using indicative and informative summaries*. PhD thesis, Columbia University, New York USA.

Kaplan, R. (1972). *The anatomy of rhetoric: Prolegomena to a functional theory of rhetoric*. Concord, MA: Heinle & Heinle.

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

Keeney, R. and Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. NY: Wiley and Sons.

Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*, chapter A Theory of Discourse Coherence. CSLI Publications, Stanford, CA.

Kennedy, L. and Ellis, D. (2003). Pitch-based emphasis detection for characterization of meeting recordings. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003)*.

Knott, A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh.

Kohavi, R. and John, G. H. (1996). Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324.

Kominek, J. and Kazman, R. (1997). Accessing multimedia through concept clustering. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 19–26, New York, NY, USA. ACM.

Koumpis, K., Renals, S., and Niranjan, M. (2001). Extractive summarization of voicemail using lexical and prosodic feature subset selection. In *Proceedings of Eurospeech*, pages 2377–2380.

Kozima, H. (1993). Text segmentation based on similarity between words. In *Proceedings of ACL 1993*.

Kreiman, J. (1982). Perception of sentence and paragraph boundaries in natural conversation . *Journal of Phonetics*, 10:163175.

Kubat, M., Holte, R., and Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3):19521.

Kunz, W. and Ritte, H. W. J. (1970). Issue as elements of information system. Technical Report Working Paper 131, Institute of Urban and Regional Development Research, University of California, Berkeley.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference*, pages 68–73.

Lafferty, J., Pereira, F., and McCallum., A. (2001). Conditional random ?elds: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*.

Langley, P. (1994). Selection of relevant features in machine learning. In *Proceedings of the AAAI 1994 Fall Symposium on Relevance*. AAAI Press.

Lehiste, I. (1980). Phonetic characteristics of discourse. In *the Meeting of the Committee on Speech Research, Acoustical Society of Japan*.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Levow, G. (2004). Prosody-based topic segmentation for mandarin broadcast news. In *Proceedings of HLT 2004*.

Levy, H. (1992). Stochastic dominance and expected utility: Survey and analysis. *MANAGEMENT SCIENCE*, 38(4):555–593.

Lewis, H. S. and Butler, T. W. (2007). An interactive framework for multi-person, multiobjective decisions. *Decision Sciences*, 24(1):1–22.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out of ACL 2004*.

Lin, C.-Y. and Hovy, E. (2003). Automatic evluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003 wORKSHOP*, pages 71–78.

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

Litman, D. and Forbes-Riley, K. (2004). Annotating student emotional states in spoken tutoring dialogues. In *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*.

Litman, D. and Passoneau, R. (1995). Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the ACL 1995*.

Litman, D. J. and Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*.

Liu, Y., Shriberg, E., Stolcke, A., and Harper, M. (2004). Using machine learning to cope with imbalanced classes in natural sppech: Evidence from sentence boundary and disfluency detection. In *Proceedings of the Intl. Conf. Spoken Language Processing*.

Lochbaum, K. E. (1991). An algorithm for plan recognition in collaborative discourse. In *Proceedings of 29th Annual Meeting of the ACL (ACL 1991)*, pages 33–38.

Maier, R. and Klosa, O. W. (1995). Organizational memory systems to support organizational information processing. *Information Systems Research*, 6(2):82–117.

Malioutov, I., Park, A., Barzilay, R., and Glass, J. (2007). Making sense of sound:unsupervised topic segmentation over acoustic input. In *Proceedings of ACL 2007*.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of Sixth Conf. on Natural Language Learning, 2002.*, pages 49–55.

Mani, I. and Bloedorn, E. (1998). Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National Conference on AI (AAAI 1998)*, pages 821–826.

Mani, I., Gates, B., and Bloedorn, E. (1999). Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the ACL.*

Mann, W. and Thompson, S. (1988). *Rhetorical structure theory: Toward a functional theory of text organization.*

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

Maskey, S. and Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Proceedings of Eurospeech*.

Matejka, P., Schwarz, P., Cernocky, J., and Chytil, P. (2005). Phonotactic language identification using high quality phoneme recognition. In *Proceedings of Eurospeech 2005*.

Matveeva, I. and Levow, G.-A. (2007). Topic segmentation with hybrid document indexing. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 351–359.

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

McKeown, K. (1985). *Text generation: using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press.

McNeill, D., editor (2000). *Language and Gesture: Window into Thought and Action*. Cambridge: Cambridge University Press.

MediaCT, I. (2008). Americas business elite embracing on-line media. http://www.ipsos-na.com/news/pressrelease.cfm?id=3922.

Menn, L. and Boyce, S. (1982). Fundamental frequency and discourse structure. *Language and Speech*, 25(

Miekes, M., Müller, C., and Strube, M. (2007). Improving extractive dialogue summarization by utilizing human feedback. In *Proceedings of AIAP 2007*, pages 627–632.

Minel, J.-L., Nugier, S., and Piat, G. (1997). How to appreciate the quality of automatic text summarization. In Mani, I. and Maybury, M., editors, *Proceedings of the ACL/EACL97 Workshop on Intelligent Scalable Text Summarization*.

Moran, T. P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., Melle, W. V., and Zellweger, P. (1997). Ill get that off the audio: A case study of salvaging multimedia meeting records. In *Proceedings of CHI 1997*, pages 202–209. ACM Press.

Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., , and Wooters, C. (2003). Meetings about meetings: research at icsi on speech in multiparty conversations. In *Proceedings of ICASP 2003*.

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

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S., Tyagi, A., Casas, J., Turmo, J., Christoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., and Rochet, C. (2008). The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Journal of Language Resources and Evaluation*, 41 (3-4):389–407.

Mosteller, F. and Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economics*, 59:371–404.

Murray, G. (2007). *Using Speech-Specific Characteristics for Automatic Speech Summarization*. PhD thesis, University of Edinburgh, Scotland, UK.

Murray, G., Renals, S., and Carletta, J. (2005). Extractive summarization of meeting recordings. In *Proceedings of Interspeech 2005*.

Murray, G., Renals, S., and Taboada, M. (2006). Prosodic correlates of rhetorical relations. In *Proceedings of HLT/NAACL 2006 ACTS Workshop*.

Nagata, M. and Morimoto, T. (1994). First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193203.

Neal, J. G. and Shapiro, S. C. (1991). *Intelligent User Interfaces*, chapter Intelligent multimedia interface technology, page 4568. ACM Press, Addison Wesley, New York.

Nenkova, A., Passonneau, R., and Kathleen McKeown, J. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transaction on Speech Language Processing*, 4(2).

Niekrasz, J., Purver, M., Dowding, J., and Peters, S. (2005). Ontology-based discourse understanding for a persistent meeting assistant. In *Proceedings of AAAI 2005 Spring Symposium*.

NIST (2002). Rich transcription 2002 stt and metadata extraction results. Technical report, RT-02 Workshop.

NIST (2004). Rich transcription 2004 spring meeting recognition evaluation. Technical report, RT-04 Workshop.

O'Keefe, D. J. (2002). *Persuasion: Theory and Research*. Thousand Oaks, CA: Sage.

Opitz, D. (1999). Feature selection for ensembles. In *Proceedings of AAAI/IAAI 1999*, pages 379–384.

Orasanu, J. and Connoly, T. (1993). *Decision making in action: Models and methods*, chapter The reinvention of decision making, pages 3–20. Norwood, NJ: Ablex Publishing Corp.

Ostendorf, M., Favre, B., Grishman, R.and Hakkani-Tur, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J., Liu, Y., Maskey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W., and Woofers, C. (2008). Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, 25(3):59–69.

O'Sullivan, J., Langford, J., Caruna, R., and Blum, A. (2000). Featureboost: A metalearning algorithm that improves model robustness. In *Proceedings of the Seventeenth International Conference on Machine Learning, 2000*, pages 703–710.

Oviatt, S., Lunsford, R., and Coulston, R. (2005). Individual differences in multimodal integration patterns: what are they and why do they exist? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 241–249.

Pallotta, V., Niekrasz, J., and Purver, M. (2005). Collaborative and argumentative models of meeting discussions. In *Workshop on Computational Models of Natural Arguments (CMNA) at the IJCAI 2005*.

Pallotta, V., Seretan, V., and Ailomaa, M. (2007a). User requirements analysis for meeting information retrieval based on query elicitation. In *Proceedings of ACL*.

Pallotta, V., Seretan, V., Ailomaa, M., Ghorbel, H., and Rajman, M. (2007b). Towards an argumentative coding scheme for annotating meeting dialogue data. In *International Conference of Pragmatics Association*.

Pang, B. and Lee, L. (2004). A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL 2004*.

Park, A. and Glass, J. R. (2006). Unsupervised word acquisition from speech using pattern discovery. in . In *Proceedings of ICASSP 2006*.

Passonneau, R. and Litman, D. (1993). Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of ACL 1993*.

Peterson, C. and Beach, L. (1967). Man as an intuitive statistician. *Psychological Bulletin*, pages 29–46.

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

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *BEHAVIORAL AND BRAIN SCIENCES*, 27:169–226.

Pietra, S. D., Pietra, V. D., and Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19:380–393.

Polanyi, L. (1988). A formal model of discourse structure. *Journal of Pragmatics*, pages 601–638.

Ponte, J. and Croft, W. (1997). Text segmentation by topic. In *Proceedings of the Conference on Research and Advanced Technology for Digital Libraries 1997*.

Popescu-Belis, A. (2004). Abstracting a dialog act tagset for meeting processing. In *Proceedings of LREC*, pages 1415–1418.

Popescu-Belis, A., Flynn, M., PierreWellner, and Baudrion, P. (2008). Task-based evaluation of meeting browsers: from task elicitation to user behavior analysis. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

Post, W. M., Cremers, A. H., and Henkemans, O. B. (2004). A research environment for meeting behavior. In *Proceedings of the 3rd Workshop on Social Intelligence Design*.

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

Quinlan, J. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman Harcourt.

Rasmussen, J. (1986). *Information processing and human-machine interaction: An approach to cognitive engineering*. NY: North Holland.

Raux, A. and Black, A. (2003). A unit selection approach to f0 modeling and its application to emphasis. In *ASRU 2003*.

Reiter, S. and Rigoll, G. (2005). Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, USA.

Reithinger, N. and Klesen, M. (1997). Dialogue act classification using language models. In *Proceedings of Eurospeech 1997*.

Renals, S. and Ellis, D. (2003). Audio information access from meeting rooms. In *Proceedings of IEEE ICASSP 2003*, pages 744–747.

Resnick, M. L. and Sanchez, J. (2004). Effects of organizational scheme and labeling on task performance in product-centered and user-centered retail websites. *Human Factors*, 46.

Reynar, J. (1998). *Topic Segmentation: Algorithms and Applications*. PhD thesis, UPenn, PA USA.

Rickert, M., Foster, M. E., Giuliani, M., By, T., Panin, G., and Knoll, A. (2007). Integrating language, vision and action for human robot dialog systems. *HCI*, 6:987–995.

Rienks, R., Heylen, D., and van der Weijden, E. (2005). Argument diagramming of meeting conversations. In *Multimodal Multiparty Meeting Processing Workshop at the ICMI*.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 105–112.

Romano, N. C. and Nunamaker, J. F. (2001). Meeting analysis: Findings from research and practice. In *Proceedings of HICSS-34*. IEEE Computer Society.

Rosenfeld, L. and Morville, P. (2002). *Information Architecture for the World Wide Web: Designing Large-scale Web Sites*. O'Reilly Media.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696735.

Saggion, H., Bontcheva, K., and Cunningham, H. (2003). Robust generic and query-based summarization. In *Proceedings of EACL 2003*, pages 235–238.

Saggion, H. and Lapalme, G. (2000). Concept identification and presentation in the context of technical text summarization. In *Proceedings of the Workshop on Automatic Summarization*.

Savage, L. (1972). *The foundations of statistics*. New York: Dover Publications.

Schafer, A. J., Speer, S., and Warren, P. (2004). *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*, chapter Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task. The MIT Press.

Schegloff, E. (1968). Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095.

Schwarz, P., Matjka, P., and ernock, J. (2004). Towards lower error rates in phoneme recognition. *Lecture Notes in Computer Science*, (3206):465–472.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University, Cambridge England.

Searle., J. R. (1990). *Collective intentionality*.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue 2004*.

Shriberg, E. and Stolcke, A. (2001). Direct modeling of prosody: An overview of applications in automatic speech processing. In *Proceedings of International Conference on Speech Prosody 2004*.

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

Simons, D. and Levin, D. (1997). Failure to detect changes to attended objects. *Invest. Opthalmol. Vis. Sci.*, 38:32733279.

Smith, P. L. and Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, 27(3):161–168.

Sparck-Jones, K. and Gallier, J. (1996). *Evaluating Natural Language Processing Systems: An analysis and Review*. Springer, Berlin.

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

Sutton, C., Sindelar, M., and McCallum., A. (2005). Feature bagging: Preventing weight undertraining in structured discriminative learning. Technical Report Technical Report IR-402., Center for Intelligent Information Retrieval, University of Massachusetts.

TDT-Evaluation (2002). The 2002 topic detection and tracking task definition and evaluation plan. Technical report, TDT2002.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Toulmin, S. (1958). *The Use of Argument*. Cambridge University Press.

Trieschnigg, D. and Kraaij, W. (2004). Tno hierarchical topic detection at tdt 2004. Technical report, TNO.

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

Trueswell, J. C. and Tanenhaus, M. K., editors (2004). *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*. The MIT Press.

Tucker, S. and Whittaker, S. (2004). Accessing multimodal meeting data: systems, problems and possibilities. In *Proceedings of MLMI 2004*.

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

Turney, P. D. (2002). Thumbs up or thumbs down: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL 2002*.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211:453–458.

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

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

Waibel, A., Bett, M., Metze, F., Ries, K., amd T. Schultz, T. S., Soltau, H., Yu, H., and Zechner, K. (2001). Advances in automatic meeting record creation and access. In *Proceedings of ICASSP 2001*.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997). Paradise: A framework for evaluating spoken dialogue agents. In Cohen, P. R. and Wahlster, W., editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Somerset, New Jersey. Association for Computational Linguistics.

Warren, P. (1999). *Language Processing*, chapter Prosody and language processing, pages 155–188. Hove: Psychology Press.

Wellner, P., Flynn, M., and Guillemot, M. (2004). Browsing recorded meetings with ferret. In *Proceedings of MLMI 2004*. Springer-Verlag.

Wellner, P., Flynn, M., Tucker, S., and Whittaker, S. (2005). A meeting browser evaluation test. In *CHI '05 extended abstracts on Human factors in computing systems*, pages 2021–2024, New York, NY, USA. ACM.

Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., Pereira, O., and Singhal, A. (1999). Scan: Designing and evaluating user interfaces to support retrieval from speech archives. In *In Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 26–33. ACM Press.

Whittaker, S., Hyland, P., and Wiley, M. (1994). Filochat: handwritten notes provide access to recorded conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 1994)*, pages 271–277, New York, NY, USA. ACM.

Whittaker, S., Laban, R., and Tucker, S. (2005). Analysing meeting records: An ethnographic study and technological implications. In *Proceedings of MLMI 2005*.

Whittaker, S., Tucker, S., Swampillai, K., and Laban, R. (2008). Design and evaluation of systems to support interaction capture and retrieval. *Personal Ubiquitous Comput.*, 12(3):197–221.

Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-05)*.

Wightman, C. W. and Talkin, D. (1994). Computational aids for the study of prosody (abstract). *Journal of the Acoustic Society of America*, 5(2).

Willard, C. A. (1988). *A Theory of Argumentation*. University of Alabama Press.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP 2005*.

Wilson, T., Wiebe, J., and Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99.

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

Wrede, B. and Shriberg, E. (2003a). The relationship between dialogue acts and hot spots in meetings. In *Proceedings of IEEE ASRU 2003 Workshop*.

Wrede, B. and Shriberg, E. (2003b). Spotting hot spots in meetings: Human judgements and prosodic cues. In *Proceedings of Eurospeech 2003*.

Wright, H. (1998). Automatic utterance type detection using suprasegmental features. In *Proceedings of ICSLP 1998*.

Xu, Z. (2008). Group decision making based on multiple types of linguistic preference relations. *Information Science*, 178(2):452–467.

Yaari, Y. (1997). Segmentation of expository texts by hierarchical agglomerative clustering. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 1997)*.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*.

Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of The Twentieth International Conference on Machine Leaning (ICML 2003)*, pages 856–863.

Zechner, K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4).

Zechner, K. and Waibel, A. (2000). DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING-2000*.

Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan., I. (2005). Semi-supervised adapted hmms for unusual event detection. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*.

Zhu, X. and Penn, G. (2005). Evaluation of sentence selection for speech summarization. In *Proceedings of RANLP 2005*.