Automatic Decision Detection in Meeting Speech

Pei-Yun Hsueh and Johanna D. Moore

School of Informatics,
2 Buccleuh Place, Edinburgh EH8 9WL, United Kingdom

Abstract. Decision making is an important aspect of meetings in organisational settings, and archives of meeting recordings constitute a valuable source of information about the decisions made. However, standard utilities such as playback and keyword search are not sufficient for locating decision points from meeting archives. In this paper, we present the AMI DecisionDetector, a system that automatically detects and highlights where the decision-related conversations are. In this paper, we apply the models developed in our previous work [1], which detects decision-related dialogue acts (DAs) from parts of the transcripts that have been manually annotated as extract-worthy, to the task of detecting decision-related DAs and topic segments directly from complete transcripts. Results show that we need to combine features extracted from multiple knowledge sources (e.g., lexical, prosodic, DA-related, and topical class) in order to yield the model with the highest precision. We have provided a quantitative account of the feature class effects. As our ultimate goal is to operate AMI DecisionDetector in a fully automatic fashion, we also investigate the impacts of using automatically generated features, for example, the 5-class DA features obtained in [2].

keywords: Spoken language understanding, meeting tracking and analysis, argumentation modelling0.

1 Introduction

Recent advances in multimedia technologies have led to huge archives of audio and video recordings of meetings. Reviewing decisions is an aspect central to our organizational life [3,4]. For example, it would be helpful for a new engineer assigned to a project to review the major decisions that have been made in previous meetings by watching the recordings. However, while it is straightforward to archive a meeting, finding out what decisions have been made from the recording is still a challenging task. Unless all decisions are recorded in meeting minutes or annotated in the audio-video recordings, it is difficult to locate the decision points using existing browsing and playback utilities alone. Moreover, a recent study [5] has shown that even when a standard keyword search utility is provided, it is still difficult to recover information about the argumentative process in the discussion (e.g., decision points).

A. Popescu-Belis, S. Renals, and H. Bourlard (Eds.): MLMI 2007, LNCS 4892, pp. 168–179, 2007.

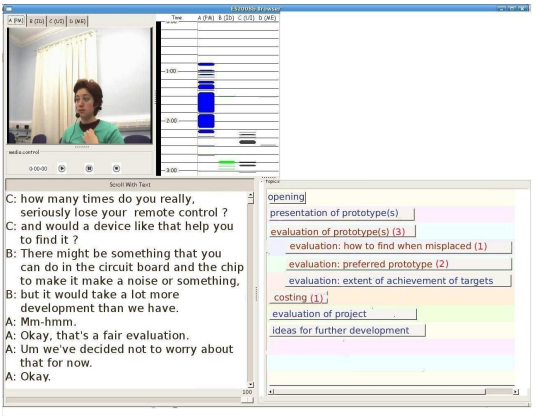

Fig. 1. Example application that demonstrates the use of decision-related topic segment information. The bottom right component shows a list of topic segments in an example meeting. The topic segments shaded in red are those that contain at least one decisions. The number shown in the parenthesis following each topic segment label indicates the number of decisions reached within the topic segment.

Banerjee and Rudnicky [6] have demonstrated that it is easier to recover information for user queries if the meeting record includes discourse-level annotations, such as topic segmentation, speaker role, and meeting state[1]. To assist users in revisiting decisions within meeting archives, our goal is thus to automatically annotate decision-related information on the dialogue acts and discussion segments where decisions are made. As the development of such an automatic decision detection component is critical to the re-use of meeting archives [7], 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) [8,9].

Previous research has developed descriptive models of meeting discussions. Some of them focus on modelling the dynamics [10], while the others focus on modelling the content [11,4]. While automatically extracting these argument models remains a challenging task, researchers have begun to make progress towards this goal [12,13,14,1,15,16].

In this paper, we present the AMI DecisionDetector, which performs automatic decision detection in meeting speech and provides visual aids for users wishing to review decisions. In particular, we are interested in locating decision-related information at two levels of granularity: topic segments and dialogue acts. First, the system detects decision-related topic segments in which meeting participants have reached at least one decision. As shown in Figure 1, this allows

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

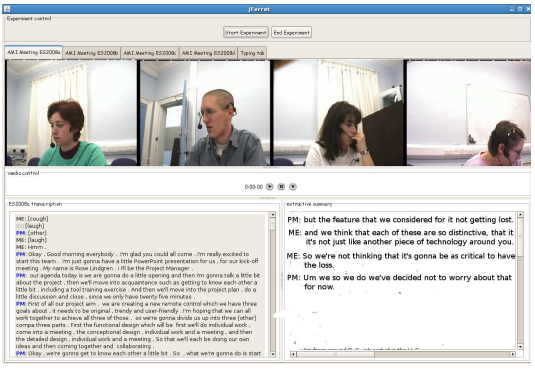

Fig. 2. Example application that demonstrates the use of decision-related DA information. The bottom right component shows a set of decision-related DA extracts that are representive of the design decision of "how to find (the remote) when misplaced".

users to get an overview of the decisions made in previous meetings by browsing the topics of the decision-related segments (e.g., those shaded in red in Figure 1).

Second, the system detects decision-related dialogue acts (DAs) by looking for DAs which are extract-worthy and reflective of the content of the decision discussions. As shown in Figure 2, this allows users to obtain details about the decisions they are particularly interested in by reviewing the relevant decision-related DAs. For example, if a user wants to know more about the design decision of "how to find (the remote) when misplaced", they can interpret the decision as "not to worry about designing a function to find the remote when misplaced" by looking at the extract shown in the bottom right component of Figure 2.

2 Data

In this study, we use a set of 50 scenario-driven meetings (approximately 37,400 DAs) that have been segmented into dialogue acts and annotated with decision information in the AMI meeting corpus [17]. These meetings are driven by a scenario, wherein four participants play the role of Project Manager, Marketing Expert, Industrial Designer, and User Interface Designer in a design team in a series of four meetings. Participants participated in only one series of 4 meetings. The corpus includes manual transcripts for all meetings as well as individual sound files recorded by close-talking microphones for each participant and cross-talking sound files recorded by an 8-element circular microphone array.

The meeting recordings have been annotated at several levels, including dialogue acts (DAs) and topics. The DA annotation scheme for the AMI corpus consists of 15 dialogue act types, which can be organised into five major groups:

- Information (31.9%): giving and eliciting information. E.g., "Suggestion".
- Action (9.8%): making or eliciting suggestions or offers. E.g., "Elicit-suggestion".

- Commenting on the discussion (22.6%): making or eliciting assessments and comments about understanding. E.g., "Assessment".
- Segmentation (31.8%): not contributing to the content but allowing segmentation of the discourse, E.g., "Backchannel", "Stall", and "Fragment".
- Other (3.9%): a remainder class for utterances which convey an intention, but do not fit into the four previous categories.

Topic segmentation and labels have also been 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 have flattened the structure into a hierarchy of two layers: top-level (TOP) and subtopic level (SUB). As the AMI meetings are scenario-driven, annotators are expected to find that most topics recur. Therefore, they are given a standard set of descriptions that can be used as labels for each identified topic 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 have collapsed these segments into one category: functional segments (FUNC). The AMI scenario meetings takes, on average, 30 minutes (around 800 DAs) and contain eight top-level topic segments and seven subtopic segments per meeting.

2.1 Decision-Related Dialogue Acts

It is difficult to determine whether a DA contains information relevant to decision without knowing what decisions have been made in the meeting. Therefore, in this study decision-related DAs are annotated in a two-phase process: First, annotators are asked to browse through the meeting record and write an abstractive summary about the decisions that have been made in the meeting. In this phase, another group of three annotators are also asked to produce extractive summaries by selecting a subset (around 10%) of DAs which form a summary of this meeting. Annotators are instructed to produce these summaries for an absent project manager.

Finally, this group of annotators are asked judge whether the DAs in the extractive summary support any of the sentences in the abstractive decision summary; if a DA is related to 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. For those extracted DAs that do not have any closely related sentence, the annotators are not obligated to specify a link. We then label the DAs that have one or more decision links as decision-related DAs.

In the 50 meetings we used for our experiments, 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 have been 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). An earlier analysis established the intercoder reliability of the

two-phase process at a kappa ranging from 0.5 to 0.8. In these experiments, for each meeting in the 50-meeting dataset we randomly choose the decision-related DA annotation of one annotator as the source of ground truth data.

2.2 Decision-Related Topic Segments

Decision-related topic segments are operationalized as the topic segments that contain one or more decision-related DAs. Overall, 198 out of 623 (31.78%) topic segments in the 50-meeting dataset are decision-related topic 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 topic segments, whereas only 7% of the Existing Product segments and none of the Trend-Watching segments are decision-related topic segments. (See Table 1 for a break-down of different types of decision-related segments.)

Table 1. Characteristics of topic segments that contain decision-related DAs

	ALL	TOP	SUB	FUNC
Percentage of Decision-related topicsegments per meeting (%)	33%	31%	35%	4%
Average number of decision-related dialogue acts per segment	3.7	4.5	2.76	3.83

3 AMI DecisionDetector

To locate decision-related information at the two levels of granularity, the AMI DecisionDetector consists of two components: (1) a decision-related DA detector which identifies important DAs pertaining to the decisions made, and (2) a decision-related topic segment detector which identifies the topic segments in which interlocutors have reached one or more decisions.

In the field of multiparty discourse understanding, previous research has commonly utilized a classification framework, in which variants of models are computed directly from data for classifying unseen instances. Models has been successfully trained for detecting the content topics [18], group activities [2,19,20], participant roles [21], addressees [22], and emotional effects (e.g., group level of interest [13], hot spots [16]). In this work, we have adopted a similar framework: the task of automatically detecting decision-related DAs is decomposed to a series of binary decisions [1]. A Maximum Entropy (MaxEnt) model is trained to automatically classify whether a DA is decision-related or not.

We evaluate the decision-related DA detector with a five-fold cross validation procedure using the set of 50 scenario-driven meetings. In each fold, a Maximum Entropy (MaxEnt) classifier is used to train models that can classify decision-related DAs on a subset of 40 meetings; next, the trained model is tested on the remaining 10 meetings that are unseen in the training phase. The decision-related

topic segment detector leverages the set of outputs (i.e., binary decisions) from the decision-related DA detector to classify whether an unseen topic segment contains any decisions. The task of detecting decision-related topic segments thus can be viewed as that of recognizing decision-related DAs in a wider window.

3.1 Features Used

Previous work has shown that combining multiple knowledge sources (e.g., words, audio-video recordings, speaker intention) is important to automatically identifying different aspects of the argumentative process [18]. For example, paralinguistic features (e.g., prosody and the amount of disfluency) have been applied to detect deceptive speech [23]. Paralinguistic features have also been combined with features that indicate speaker intention (i.e., DA classes) to detect "hot spots"[2] [24,16]. Similarly, lexical features, such as occurrence counts of cue words, have been used to detect learning attitudes of students in a tutoring system [25] and to detect where speakers are agreeing with one another [12,14].

Here we are interested in examining the merits of multimodal feature combinations on the performance of AMI DecisionDetector. In particular, we examine the use of the following features:

Prosodic Features: Our previous work [1] has shown that there exist prominent acoustic characteristics of decision-related DAs. For example, when it comes to a decision point, interlocutors either speak very fast or very slowly; the pitch usually goes up first and then goes down in the midpoint of a dialogue act. In this work, we use the same set of prosodic features, i.e., duration, pause, speech rate, pitch contour, and energy level. For details of how to generate these features with Shriberg and Stolcke's direct modelling approach [26], please refer to [27]. An exploratory study has shown the benefits of including immediate prosodic contexts, and thus we also include prosodic features of the immediately preceding and following DA.

Lexical Features: Previous work has also shown the importance of the lexical characteristics of decision-related DAs. For example, interlocutors use "*We*" more than "*I*" and "*You*" when reaching a decision. Likewise, they also explicitly mention topical words, such as "*advanced chips*" and "*functional design*", and use fewer negative expressions, such as "*I don't think*" and "*I don't know*". Thus we also include lexical items in our feature sets. In each fold of cross validation, we compile a list of cue words, which have occurred more than once in the set of decision-related DAs in the "training" set of meetings. Each DA is then represented as a vector of unigrams in the list of cue words.

DA-related Features: These include DA classes and speaker roles (e.g., project manager, marketing expert). We also include DA classes of the immediately preceding and following DA. As mentioned in Section 2, we have grouped the 15 DA classes (15-Class) into five major groups (5-Class). We have also obtained the automatic 5-Class predictions for each DA [2]. The accuracy of the automatic DA

[2] Namely locations that exhibit a high level of affect in the voices of interlocutors.

class predictions is 59.1%. In the following experiment, we thus can evaluate the impact of the three different versions of DA class information: manual 15-Class, manual 5-Class, and automatic 5-Class.

Topical Features: As reported in Section 2.2, we find that interlocutors are more likely to reach decisions when certain topics are brought up. Also, we expect decision-making conversations to take place towards the end of a topic segment. Therefore, we include the following features: the label of the current topic segment, the position of the DA in a topic segment (measured in words, in seconds, and in %), the distance to the previous topic shift (both at the top-level and subtopic level)(measured in seconds), the duration of the current topic segment (both at the top-level and subtopic level)(measured in seconds).

4 Results

In Section 3.1, we described the four major types of features used in this study: prosodic (PROS), unigrams (LX1), DA-related (DA), and topical (TOPIC) features. As opposed to our previous work, which detects decision-related DAs on only the parts of meetings that have been identified as extract-worthy, we trained models to detect decision-related DAs directly from entire transcripts. 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 has gone from 14% down to 2%. For comparison, we use the lexical models trained with the unigram lexical features (LX1) as our baseline.[3] The different combinations of features we used for training models can be divided into the three groups: (A) lexical features alone (BASELINE); (B) all available features except one of the four types of features (ALL-LX1, ALL-PROS, ALL-DA, ALL-TOPIC); and (C) all available features (ALL).

4.1 Experiment 1: Classifying Decision-Related Dialogue Acts and Topic Segments

Table 2 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-related DAs, the results in Table 2 are obtained using a lenient match measure, allowing a window of 20 seconds preceding and following a hypothesized decision-related DA for recognition. The task of detecting decision-related topic segments can be viewed as that of detecting decision-related DAs in a wider window. The right most three columns of the training set and test set results in Table 2 show the results of detecting decision-related topic segments.

The results demonstrate that, compared to the LX1 baseline, models trained with all features (ALL), including lexical, prosodic, DA-related and topical features, yield notably better precision on the task of decision-related topic segment

[3] 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.

Table 2. Effects of different combinations of features on detecting decision-related DAs and topic segments

	TRAIN SET						TEST SET					
Decision-Related	Dialogue Act			Topic Segment			Dialogue Act			Topic Segment		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BASELINE(LX1)	0.32	0.47	0.38	0.44	0.65	0.52	0.22	0.39	0.28	0.39	0.65	0.49
ALL-LX1	0.64	0.10	0.18	0.72	0.13	0.22	0.35	0.07	0.11	0.52	0.17	0.25
ALL-PROS	0.69	0.46	0.55	0.72	0.53	0.61	0.30	0.16	0.21	0.51	0.38	0.43
ALL-DA	0.70	0.48	0.57	0.72	0.56	0.63	0.32	0.24	0.26	0.49	0.44	0.46
ALL-TOPIC	0.64	0.36	0.46	0.70	0.48	0.57	0.24	0.18	0.20	0.49	0.41	0.44
ALL	0.72	0.38	0.49	0.74	0.45	0.55	0.35	0.19	0.24	0.56	0.38	0.44

prediction, 74% on the training set and 56% on the test set. However, in the test set, the overall accuracy (F1 score) of the combined models is relatively worse than the baseline, due to the substantially lower recall rate.

To study the relative effect of the different feature types, Rows 2-5 in the table report the performance of models in Group B, which are trained with all available features except LX1, PROS, DA and TOPIC, respectively. The amount of degradation in the overall 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-Assessment", 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-related DAs.

In sum, results suggest that (1) lexical features are the most predictive in terms of overall accuracy, despite low precision, (2) prosodic features have positive impacts on precision but not on recall, and (3) DA-related and topical features are both beneficial to precision but detrimental to recall.

4.2 Experiment 2: Exploring Automatically Generated DA Class Features

As our ultimate goal is to operate AMI DecisionDetector in an automatic fashion, we evaluate the impact of the automatically generated DA class features on the task of detecting decision-related DAs and topic segments. We have utilized the 5-class DA predictions (AUTO-5DA) generated in [2]. To understand whether the automatically generated features caused any degradation, we train models which combine all available lexical, prosodic and topical features with the AUTO-5DA features. We then evaluate the AUTO-5DA model against other models

Table 3. Effects of different versions of DA class features on detecting decision-related DAs and topic segments

	TRAIN SET						TEST SET					
Decision-Related	Dialogue Act			Topic Segment			Dialogue Act			Topic Segment		
Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EXTRACT (MANUAL-15DA)	0.73	0.61	0.66	0.74	0.66	0.69	0.37	0.38	0.36	0.50	0.53	0.50
EXTRACT (MANUAL-5DA)	0.70	0.77	0.70	0.68	0.73	0.70	0.36	0.44	0.39	0.49	0.62	0.54
EXTRACT (AUTO-5DA)	0.69	0.71	0.70	0.68	0.73	0.70	0.33	0.40	0.35	0.46	0.55	0.49

which combine the other features with the two types of manually annotated dialogue act class features: MANUAL-5DA and MANUAL-15DA. The results reported here are obtained by operating AMI DecisionDetector on the part of meetings that have been manually annotated as extract-worthy. This is because we want to focus on analyzing the impacts of the automatic DA features on the task of decision detection, rather than on that of extractive summarization.

Results in Table 3 show that our strategy that groups 15 DA classes into five major classes is beneficial to the models on the task of decision detection. It improves the recall of predicting decision-related topic segments by 18%. Replacing the manual 5-class DA features with the automatically generated version degrades the performance by 10%. However, the accuracy of prediction using the 5 automatically predicted DA classes (AUTO-5DA) compares favorably with the accuracy when using the 15 manually annotated DA classes (MANUAL-15DA).

5 Conclusions and Future Work

In this paper, we present AMI DecisionDetector, a system which performs automatic decision detection in meeting speech and provides visual aids for users who wish to review decisions. To avoid the costly requirement of operating on extractive summaries, we have examined how our computational models perform when detecting decisions directly from complete meeting transcripts. We have evaluated the models on the task of predicting decision-related discussions at two levels of granularity: dialogue acts and topic segments. To further overcome the problem of imbalanced class distribution (i.e., only 2% are positive cases), we have leveraged a variety of knowledge sources (e.g., words, prosody, DA-related contexts, topic annotations). Experimental results suggest that the model combining all the available knowledge sources performs substantially better, achieving 74% and 56% precision on the task of detecting decision-related topic segments in the training set and test set respectively. The framework we applied here can also be used to recover information for other aspects in the argumentation process, such as problems and action items.

We have also provided a quantitative account of the merits of different feature classes. Among features that are extracted from the widely ranging knowledge sources, lexical features are the most indispensable. Also, DA-related features can improve the precision of models but degrade the recall. These findings are

consistent with the results of our previous experiment which operates AMI DecisionDetector on a selective set of dialogue acts in the transcripts.

However, there are also other findings that no longer hold true when our system is operated on complete transcripts. For example, [1] has shown topical features have a distinctive advantage for locating decision topic segments from extractive summaries. However, this is not the case when identifying decision points in entire transcripts. In addition, the model trained with lexical features alone outperforms the combined model in its recall rate. This is possibly because when attempting to detect decisions from the whole transcripts, the system needs to simultaneously disambiguate the extract-worthy and decision-related dialogue acts. Therefore, features that are good at disambiguating both will stand out, and features that fail in the extract-worthy DA detection task will be shown as weak features to the final performance of decision-related DA detection.

Another drawback of our previous approach is that many of the features used in this study require human intervention, such as manual transcriptions, annotated DA segmentations and labels, and other types of meeting-specific features (e.g., speaker role). However, these semi-automatic and manual features are not always available. Therefore, in this work we tested whether our system is robust to the noise introduced by the automatically generated versions of these features. An exploratory study has shown that the performance of our approach does not degrade considerably after replacing the reference words with the ASR words, despite word recognition errors. Our further investigation on the impacts of using an automatically generated version of the DA class features (as reported in [2]) shows that it is possible to include these automatic features in the model directly. It will not degrade the performance more than including the manually annotated 15-class DA features in the first place.

Also, our approach which automatically extracts decision-related DAs as summaries has some liabilities. First, the unconnected DAs in the extract result in semantic gaps that require contextualization to bridge. Second, anaphora and unexpected topic shifts between these extracted DAs also require context to resolve. Previously, we have attempted to provide such contexts by indicating the topic of the current discussion. However, a preliminary study has shown that the segment boundaries of decision-related discussions coincide with that of the topic segments less than 50% of the time. 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 more efficiently and effectively, this assumption has yet to be tested with human subjects.

Therefore, we are now planning to conduct an extrinsic decision audit task-based evaluation on the utility of displaying decision-related DA information (as exemplified in Figure 2) to the users. We have also annotated decision-related discussion segmentation, which can be used to train computational models to find contexts that are needed for the interpretation of the identified decision points. Moreover, as we would like to disambiguate which sentence in the abstractive decision summary of a meeting is the most relevant to each of the identified decision points, the decision discussion segmentation annotations can also form a foundation for the development of the disambiguation model.

Acknowledgments

This work was supported by the European Union 6th FWP IST Integrated Project FP6-033812 AMIDA (Augmented Multiparty Interaction with Distance Access). We thank Alfred Dielmann for generating the 5-Class DA predictions for us, Jonathan Kilgour and Jean Carletta for their continuous support on the development of the AMI DecisionDetector system we demoed in MLMI 2007 at Brno, Czech Republic, and Theresa Wilson for her insightful feedbacks on the decision discussion segmentation annotation work. We also thank Steve Renals and the three anonymous reviewers for their enlightening comments on this paper.

References

1. Hsueh, P., Moore, J.: What decisions have you made: Automatic decision detection in conversational speech. In: Proceedings of NACCL/HLT 2007 (2007)
2. Dielmann, A., Renals, S.: DBN based joint dialogue act recognition of multiparty meetings. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (2007)
3. Pallotta, V., Niekrasz, J., Purver, M.: Collaborative and argumentative models of meeting discussions. In: IJCAI 2005. Proceeding of CMNA-05 workshop on Computational Models of Natural Arguments (2005)
4. Rienks, R., Heylen, D., van der Weijden, E.: Argument diagramming of meeting conversations. In: Multimodal Multiparty Meeting Processing Workshop at the ICMI (2005)
5. Pallotta, V., Seretan, V., Ailomaa, M.: User requirements analysis for meeting information retrieval based on query elicitation. In: Proceedings of ACL 2007 (2007)
6. Banerjee, S., Rose, C., Rudnicky, A.I.: The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In: Proc. of the International Conference on Human-Computer Interaction (2005)
7. Whittaker, S., Laban, R., Tucker, S.: Analysing meeting records: An ethnographic study and technological implications. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, Springer, Heidelberg (2006)
8. Post, W.M., Cremers, A.H., Henkemans, O.B.: A research environment for meeting behavior. In: Proceedings of the 3rd Workshop on Social Intelligence Design (2004)
9. Romano, N.C., Nunamaker, J.F.: Meeting analysis: Findings from research and practice. In: Proceedings of HICSS-34, IEEE Computer Society Press, Los Alamitos (2001)
10. Niekrasz, J., Purver, M., Dowding, J., Peters, S.: Ontology-based discourse understanding for a persistent meeting assistant. In: Proc. of the AAAI Spring Symposium (2005)
11. Marchand-Mailet, S.: Meeting record modeling for enhanced browsing. Technical report, Computer Vision and Multimedia Lab, Computer Centre, University of Geneva, Switzerland (2003)
12. Galley, M., McKeown, J., Hirschberg, J., Shriberg, E.: Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In: Proc. of the 42nd Annual Meeting of the ACL (2004)
13. Gatica-Perez, D., McCowan, I., Zhang, D., Bengio, S.: Detecting group interest level in meetings. In: ICASSP 2005. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, IEEE Computer Society Press, Los Alamitos (2005)

14. Hillard, D., Ostendorf, M., Shriberg, E.: Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In: Proc. HLT-NAACL 2003 (2003)
15. Purver, M., Ehlen, P., Niekrasz, J.: Shallow discourse structure for action item detection. In: The Workshop of HLT-NAACL: Analyzing Conversations in Text and Speech, ACM Press, New York (2006)
16. Wrede, B., Shriberg, E.: Spotting hot spots in meetings: Human judgements and prosodic cues. In: Proceedings of EUROSPEECH 2003 (2003)
17. Carletta, J., et al.: The AMI meeting corpus: A pre-announcement. In: Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (2005)
18. Hsueh, P., Moore, J.D.: Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In: Proceedings of the 45th Annual Meeting of the ACL (2007)
19. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic analysis of multimodal group actions in meetings. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27(3), 305–317 (2005)
20. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan., I.: Semi-supervised adapted hmms for unusual event detection. In: CVPR 2005. Proc. IEEE International Conference on Computer Vision and Pattern Recognition, IEEE Computer Society Press, Los Alamitos (2005)
21. Banerjee, S., Rudnicky, A.I.: You are what you say: Using meeting participants' speech to detect their roles and expertise. In: Proceedings of the Workshop of HLT-NAACL: Analyzing Conversations in Text and Speech, ACM Press, New York (2006)
22. Jovanovic, N., op den Akker, R., Nijholt, A.: Addressee identification in face-to-face meetings. In: The Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
23. Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., Kajarekar, S.: Combining prosodic, lexical and cepstral systems for deceptive speech detection. In: Proc. IEEE ICASSP, IEEE Computer Society Press, Los Alamitos (2006)
24. Wrede, B., Shriberg, E.: The relationship between dialogue acts and hot spots in meetings. In: Proceedings of IEEE ASRU Workshop, IEEE Computer Society Press, Los Alamitos (2003)
25. Litman, D.J., Forbes-Riley, K.: Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Communication (2006)
26. Shriberg, E., Stolcke, A.: Direct modeling of prosody: An overview of applications in automatic speech processing. In: International Conference on Speech Prosody 2004 (2001)
27. Murray, G., Renals, S., Taboada, M.: Prosodic correlates of rhetorical relations. In: Proceedings of HLT/NAACL ACTS Workshop (2006)