What Decisions Have You Made?: Automatic Decision Detection in Meeting Conversations

Abstract

This study addresses the problem of automatically detecting decisions in conversational speech. We formulate the problem as classifying decision-making units at two levels of granularity: dialogue acts and topic segments. We conduct an empirical analysis to determine the characteristic features of decision-making dialogue acts, and train MaxEnt models using these features for the classification tasks. We find that models that combine lexical, prosodic, and topical features yield the best results on both tasks, achieving 64% and 83% overall accuracy, respectively. The study also provides a quantitative analysis of the relative importance of the feature types.

1 Introduction

Making decisions is an important aspect of conversations in collaborative work. In the context of meetings, the proposed argumentative models, e.g., in Pallotta et al. (2005) and Rienks et al. (2005), have specified decisions as an essential outcome of meetings. Whittaker et al. (2005) have also described use cases in which reviewing decisions is critical to the re-use of meeting recordings. For example, when an engineer is assigned to an ongoing project, he will need to find out what major decisions have been made in previous meetings. Unless all decisions are recorded in meeting minutes or annotated in the speech recordings, it is difficult to locate the decision points using browsing and playback utilities alone.

Banerjee and Rudnicky (2005) have shown that it is easier for users to retrieve the information they seek if the meeting record includes information about topic segmentation, speaker role, and meeting state (e.g., discussion, presentation, briefing). To assist users in identifying or revisiting decisions in meeting archives, our goal is to automatically identify the dialogue acts and segments where decisions are made. Because reviewing decisions is indispensable in collaborative work, automatic decision detection is expected to lend support to computer-assisted meeting tracking and understanding (e.g., assisting in the fulfilment of the decisions made in the meetings) and the development of group information management applications (e.g., constructing group memory).

2 Related Work

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), and human-computer dialogues (e.g., DARPA Communicator). In order to develop techniques for understanding multiparty meeting dialogues, instrumented meeting rooms have been built at several institutes to record large corpora of meetings in natural contexts, including CMU (Waibel et al., 2001), LDC (Cieri et al., 2002), NIST (Garofolo et al., 2004), ICSI (Janin et al., 2003), and in the context of the IM2/M4 project (Marchand-Maillet,

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 project (the Y2 Scenario Data) (CALO, 2003) and the AMI project (Carletta et al., 2005).

The availability of meeting corpora has enabled researchers to begin to develop descriptive models of meeting discussions. Some researchers are attempting to model 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 interested in modelling the content of the meeting using the type of structures that have been 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-Maillet (2003) proposed a schema to model different argumentation acts (e.g., accept, request, reject) and their organization and synchronization. Decisions are often seen as a by-product of these models.

Automatically extracting these argument models is a challenging task. However, researchers have begun to make progress towards this goal. For example, Gatica et al. (2005) and Wrede and Shriberg (2003) 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 (N-grams) 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 automatically 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 multiparty meetings. Moreover, no previous research has provided a quantitative account of the effects of different feature types on the task of automatic decision detection.

3 Research Goal

Our aim is to develop models for automatically detecting segments of conversation that contain decisions, and to identify the feature combinations that are most effective for this task.

Meetings can be viewed at different levels of granularity. In this study, we first consider how to detect the dialogue acts that contain decision-related information (DM DAs). Since it is often difficult to interpret a decision without knowing the current topic of discussion, we are also interested in detecting decision-making segments at a coarser level of granularity: topic segments. The task of automatic decision detection can therefore be divided into two subtasks: detecting DM DAs and detecting decision-making topic segments (DM Segments).

In this study we first empirically identify the features that are most characteristic of decision-making dialogue acts and then computationally integrate the characteristic features to locate the DM DAs in meeting archives. We then use a Maximum Entropy (MaxEnt) classifier to combine the decision-characteristic features to predict DM DAs and DM Segments.

4 Data

4.1 Decision Annotation

In this study, we use a set of 50 scenario-driven meetings (approximately 37,400 dialogue acts) that have been segmented into dialogue acts and annotated with decision information in the XXX meeting corpus. These meetings are driven by a scenario, wherein four participants play the role of Project Manager (PM), Marketing Expert (ME), Industrial Designer (ID), and User Interface Designer (UID) in a design team in a series of four meetings. The corpus includes manual transcripts for all meetings. It also comes with individual sound files recorded by close-talking far field head-mounted microphones and cross-talking sound files recorded by desktop microphones.

4.1.1 Decision-Making Dialogue Acts

DM DAs are annotated in a two-phase process: First, annotators are asked to browse through the meeting record and write an abstractive summary directed to the project manager about the decisions that have been made in the meeting. Next, another group of annotators are asked to produce extractive summaries by selecting a subset (around 10%) of dialogue acts which form a summary of this meeting for the absent manager to understand what has transpired in the meeting.

Finally, this group of annotators are asked to go through the extractive dialogue acts one by one and judge whether they support any of the sentences in the decision section of the abstractive summary; if a dialogue act is related to any sentence in the decision section, a "decision link" from the dialogue act to the decision sentence is added. For those extracted dialogue acts that do not have any closely related sentence, the annotators are not obligated to specify a link. We then label the dialogue acts that have one or more decision links as DM DAs. In the 50 meetings we used for the experiments, 554 out of 37,400 dialogue acts have been annotated as DM DAs, accounting for 1.4% of all dialogue acts in the data set and 12.7% of the extracted dialogue acts.

4.1.2 Decision-Making Topic Segments

Topic segmentation has also been annotated for the XXX meeting corpus. Annotators had the freedom to mark a topic as subordinated (down to two levels) wherever appropriate. Because the XXX meetings are scenario-driven, annotators are expected to find that most topics recur. Therefore, they are given a standard set of topic descriptions that can be used as labels for each identified topic segment. Annotators will only add a new label if they cannot find a match in the standard set.

DM Segments are operationalized as topic segments that contain one or more DM DAs. Overall, 198 out of 623 (31.78%) topic segments in the 50-meeting dataset are DM Segments. Because the meetings are driven by a predetermined agenda, we expect that certain topics are more likely to involve decision making that others, and this is indeed what our analysis shows. For example, 80% of the segments labelled as Costing and 58% of those labelled Budget are DM Segments, whereas only 7% of the Existing Product segments and none of the Trend-Watching segments are DM Segments. Functional segments, such as Chitchat, Opening and Closing, almost never include decisions.

4.2 Features Used

To provide a qualitative account of the effect of different feature types on the task of automatic decision detection, we have conducted empirical analysis on four major types of features: lexical, prosodic, contextual and topical features.

4.2.1 Lexical Features

Previous research has studied lexical differences (i.e., occurrence counts of N-grams) in various aspects of spoken language, including speaker gender (Boulis and Ostendorf, 2005), story-telling (Gordon and Ganesan, 2005), and topics (Anonymous and Anonymous, 2006). Because we expect that lexical differences will also distinguish DM units from non-DM units in meeting conversations, we generated language models for these two groups of utterances in our corpus. Comparison of the language model generated from the DM DAs with the model trained from the remaining, non-DM DAs shows that there are indeed differences: (1) decision making DAs are more likely to contain *we* than *I* and *You*; (2) in DM DAs there are more explicit mentions of topical words, such as *advanced chips* and *functional design*; (3) in DM DAs, there are fewer negative expressions, such as *I don't think* and *I don't know*. In an exploratory study using unigrams, bigrams and trigrams, we found that using bigrams and trigrams does not improve the accuracy of classifying DM DAs, and therefore we include only unigrams in the set of lexical features in the experiments reported in Section 6.

4.2.2 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., energy, and fundamental frequency (F0), are indicative of segmentation and saliency. In this study, we follow Shriberg and Stolcke's (2001) direct modelling approach to manifest prosodic features as duration, pause, speech rate, pitch contour, and energy level. We utilize the individual sound files pro-

vided in the XXX corpus. To extract prosodic features from the sound files, we use the Snack Sound Toolkit to compute a list of pitch and energy values delimited by frames of 10 ms, using the normalized cross correlation function. Then we apply 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 of a dialogue act is approximated by measuring the pitch slope at multiple points within the dialogue act, e.g., the first and last 100 and 200 ms. The rate of speech is calculated as both the number of words spoken per second and the number of syllables per second. We use Festival's speech synthesis front-end to return phonemes and syllabification information. Prior work has shown the benefits of including immediate prosodic contexts, and thus we also include prosodic features of the immediately preceding and following dialogue acts. Table 1 contains a list of automatically generated prosodic features used in this study.

4.2.3 Contextual Features

From our qualitative analysis, we expect that contextual features specific to the XXX corpus, such as speaker role (i.e., PM, ME, ID, UID) and meeting type (i.e., kick-off, conceptual design, functional design, detailed design) will be characteristic of the DM DAs. Indeed, our analysis shows that (1) participants assigned to the role of PM produce 42.5% of the DM DAs, and (2) participants make relatively fewer decisions in the kick-off meetings. Moreover, we found that some types of DAs are more likely to be DM DAs than others. In particular, dialogue acts of type ***inform***, ***suggest***, ***elicit assessment***, and ***elicit inform*** are more likely to be DM DAs than other dialogue act types in the annotation scheme.

We have also found that immediately preceding and following dialogue acts are important for identifying DM DAs. For example, ***stalls*** and ***fragments*** preceding and ***fragments*** following a DM DA are more likely than for non-DM DAs.[1] In contrast, there is a lower chance of seeing a ***suggest*** act or an eliciting act (i.e., ***elicit-inform***, ***elicit-suggestion***, ***elicit-assessment***) in the preceding and following DM DAs.

[1] A dialogue act is marked as a ***stall*** when a speaker starts talking before they are ready, or keeps speaking when they haven't figured out what to say, just to try and keep the attention of the group. An act is marked as a ***fragment*** if the speaker did not get far enough to express an intention, or stopped altogether, or is a unit that is not clear enough to be transcribed.

4.2.4 Topical Features

As reported in Section 4.1.2, we find that interlocutors are more likely to reach decisions when certain topics are brought up. We also expect decisions to take place towards the end of a topic segment. Therefore, in our study we included the following topic 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 of the DA from the previous topic shift (both at the top-level and sub-topic level)(measured in seconds), and the duration of the current topic segment (both at the top-level and sub-topic level)(measured in seconds).

5 Experiment

5.1 Classifying DM DAs

Detecting DM DAs is the first step of automatic decision detection. For this purpose, we trained MaxEnt models to classify each unseen sample as either DM DA (POS) or non-DM DA (NEG). We performed a 5-fold cross validation on the set of 50 meetings. In each fold, we trained 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 were used to classify unseen instances in the test set as either POS or NEG. In Section 4.2, we described the four major types of features used in this study: unigrams (LX1), prosodic (PROS), contextual (CONT), 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 generated fully automatically. 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 each of lexical, contextual and topical features alone (LX1, CONT, TOPIC); (C) using all available features except one of the four types of features (ALL-LX1, ALL-PROS, ALL-CONT, ALL-TOPIC); and (D) using all available features (ALL).

Type	Feature
Duration	Number of words spoken in current, previous and next dialogue act
	Duration (in seconds) of current, previous and next dialogue act
Pause	Amount of silence (in seconds) preceding a dialogue act
	Amount of silence (in seconds) following a dialogue act
Speech rate	Number of words spoken per second in current, previous and next dialogue act
	Number of syllables per second in current, previous and next dialogue act
Energy	Overall energy level
	Average energy level in the first, second, third, and fourth quarter of a dialogue act
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 dialogue act

Table 1: *Prosodic features used in this study.*

5.2 Classifying DM Segments

Detecting DM segments is necessary for interpreting decisions, because the segment provides information about the current topic of discussion. Here we combine the predictions of the DM DAs to classify each unseen topic segment in the test set as either DM Segment (POS) or non-DM Segment (NEG). Recall that we defined a DM Segment as a segment that contains one or more hypothesized DM DAs. The task of detecting DM Segments can thus be viewed as that of detecting DM Dialogue Acts in a wider window.

6 Results

6.1 Experiment 1: Classifying DM DAs

Table 2 reports the performance on the test set. The last three columns are the results obtained using a lenient match measure, allowing a window of 10 seconds preceding and following a hypothesized DM DA for recognition.

The results show that models trained with all available features (ALL), including lexical, prosodic, contextual and topical features, yield substantially better performance than the baseline on the task of detecting DM DAs. We carried out a one-way ANOVA to examine the effect of different feature combinations on overall accuracy (F1). The ANOVA suggests a reliable effect of feature type ($F(9, 286) = 3.44; p < 0.001$). Rows 2-4 in Table 2 report the performance of models in Group B that are trained with a single type of feature. We see that lexical features are the most predictive features when used alone. We performed sign tests to determine whether there are statistical differences among these models and the baseline. We find that when used alone, only lexical features (LX1) can train a better model than the baseline ($p < 0.001$). However, none of these models yields performance comparable to the ALL model.

To study the relative effect of the different feature types, Rows 5-8 in Table 2 report the performance of models in Group C, which are trained with all available features except LX1, PROS, CONT and TOPIC features, 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 of the model. We performed sign tests to examine the differences among these models and the ALL model. We find that the ALL model outperforms all of these models ($p < 0.001$) except the model trained by leaving out contextual features (ALL-CONT). A closer investigation of the precision and recall of the ALL-CONT model shows that the contextual features are beneficial for precision but detrimental to recall. The mixed result is due to the fact that models trained with contextual features are tailored to recognize particular types of DM DAs. Therefore, using these contextual features improves the precision for these types of DM DAs but reduces the overall recognition accuracy.

	Exact Match			Lenient Match		
Accuracy	Precision	Recall	F1	Precision	Recall	F1
BASELINE(PROS)	0.32	0.06	0.1	0.32	0.1	0.15
LX1	0.53	0.30	0.38	0.6	0.43	0.5
CONT	0	0	0	0	0	0
TOPIC	0.49	0.11	0.17	0.57	0.11	0.17
ALL-PROS	0.63	0.47	0.54	0.71	0.57	0.63
ALL-LX1	0.61	0.34	0.44	0.65	0.43	0.52
ALL-CONT	0.66	0.62	**0.64**	0.69	0.68	0.69
ALL-TOPIC	0.72	0.54	0.62	0.70	0.52	0.59
ALL	0.72	0.54	0.62	0.76	0.64	0.70

Table 2: *Effects of different combinations of features on detecting DM DAs.*

6.2 Experiment 2: Classifying DM Segments

As expected, the scores reported in Table 3 are higher than those reported in Table 2, with best performance of 83% overall accuracy. The model that combines all available features (ALL) yields significantly better results than the baseline. An ANOVA shows a reliable effect of different feature types on the task of detecting DM Segments ($F(11, 284) = 2.33; p <= 0.01$). Rows 2-4 suggest that lexical features are the most predictive features in terms of overall accuracy. Sign tests confirm the advantage of using lexical features (LX1) over the baseline (PROS) ($p < 0.05$). Interestingly, the model that is trained with topical features alone (TOPIC) yields substantially better precision ($p < 0.001$). The increase from 49% precision for the task of detecting DM DAs (see Table 2) to 91% for that of detecting DM Segments stems from the fact that decisions are more likely to occur in certain types of topic segments. In turn, training models with topical features helps eliminate incorrect predictions of DM DAs in these types of topic segments. However, the high accuracy gained by using the TOPIC model on detecting certain types of DM Segments does not extend to all types of DM Segments. This is shown by the significantly lower recall of the TOPIC model over the baseline ($p < 0.001$).

Finally, Rows 5-8 report the performance of the models in Group (C) on the task of detecting DM Segments. Sign tests again show that the model that is trained with all available features (ALL) outperforms the models that leave out lexical, prosodic, or topical features ($p < 0.05$). However, once again, the ALL model does not outperform the model that leaves out contextual features. This is because the inclusion of contextual features improves precision, but degrades recall for this task just as it did for the DM DA detection task. This is not surprising since we have defined DM segments as those segments containing one or more DM DA. For the other three feature types, calculating how much the overall 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 topical and prosodic features.

	Exact Match		
Accuracy	Precision	Recall	F1
BASELINE(PROS)	0.67	0.39	0.49
LX1	0.69	0.69	0.69
CONT	0	0	0
TOPIC	0.91	0.17	0.29
ALL-PROS	0.82	0.76	0.79
ALL-LX1	0.79	0.64	0.70
ALL-CONT	0.79	0.86	**0.83**
ALL-TOPIC	0.75	0.73	0.74
ALL	0.86	0.80	0.82

Table 3: *Effects of different combinations of features on detecting DM Segments.*

7 Discussion

As suggested by the mixed results obtained by the model that is trained without contextual features, the two-phase decision annotation procedure (as described in Section 4.1) may have caused annotators

(1) A: but um the feature that we considered for it not getting lost.
(2) B: Right. Well
(3) B: were 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 1: *Example decision-making discussion*

to select dialogue acts that serve different functional roles in a decision-making process in the set of DM DAs. For example, in the dialogue shown in Figure 1, the annotators have marked dialogue acts (1), (5), (8), and (11) as the DM DAs related to this decision: *"There will be no feature to help find the remote when it is misplaced"*. Among the four DM DAs, (1) describes the topic of what this decision is about; (5) and (8) describe the arguments that support the decision-making process; (11) indicates the level of agreement or disagreement for this decision. Yet these DM DAs play different functional roles in the decision making process, and may thus each have their own characteristic features. Training one model to recognize DM DAs of all functional roles may have degraded the performance on the classification tasks. Developing models for detecting DM DAs that play different functional roles requires a larger scale study to discover the "anatomy" of decision-making discussions.

8 Conclusions and Future Work

This is the first study that aimed at detecting segments of meeting conversation that contain decisions. We have (1) empirically analyzed the characteristic features of DM dialogue acts, and (2) trained models to detect DM dialogue acts and DM topic segments, given the set of characteristic features. Empirical analysis has provided a qualitative account of the DM-characteristic features, whereas training the computational models on different feature combinations has provided a quantitative account of the effect of different feature types on the task of automatic decision detection. Empirical analysis has exhibited demonstrable differences in the lexical items (e.g., *we*), the contextual features (e.g., *meeting type*, *speaker role*, *dialogue act type*), and the topical features associated with DM DAs. The experimental results have suggested that (1) the model combining lexical, prosodic, and topical features performs substantially better than other models, achieving 64% and 83% overall accuracy on the task of detecting DM DAs and that of detecting DM Segments, respectively, (2) the contextual features we employed in these studies improve precision, but degrade recall, and thus degrade overall accuracy, (3) lexical features are the best indicators for both the task of detecting DM DAs and that of detecting DM Segments, and (4) combining topical features is important for improving precision for the task of detecting DM Segments.

Many of the features used in this study were produced by human transcribers or annotators. This includes manual transcription, annotation of dialogue act segmentation and labels, annotation of topic segmentation and labels, and other types of meeting-specific features. Our ultimate goal is to identify decisions using automatically induced features. Therefore, studying how performance degrades when we use automatically generated versions of these features is essential for developing a fully automated component for detecting decisions immediately after a meeting or while a meeting is still in progress. In addition, as pointed out in Sections 6 and 7, DM DAs play different functional roles in the decision making process, but have all been annotated as DM DAs in the current annotation. Purver et al. (2006) have suggested a hierarchical annotation scheme to

accommodate the different aspects of action items, which are a specific type of decision made in meetings. The same technique may be applicable to decision detection more generally, and we plan to pursue this in our future work.

References

A. Anonymous and B. Anonymous. 2006. Automatic topic segmentation and labelling in multiparty dialogue. In *First IEEE/ACM workshop on Spoken Language Technology (SLT)*. IEEE/ACM.

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

C. Boulis and M. Ostendorf. 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*. ACM Press.

CALO. 2003. http://www.ai.sri.com/project/calo.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meeting corpus: A pre-announcement. In *Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*. See http://www.amiproject.org.

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

M. Galley, J. McKeown, J. Hirschberg, and E. Shriberg. 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*.

J. S. Garofolo, C. D. Laprun, M. Michel, V.M. Stanford, and E. Tabassi. 2004. The nist meeting room pilot corpus. In *Proceedings of LREC04*.

D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. 2005. Detecting group interest level in meetings. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*.

Andrew S. Gordon and Kavita Ganesan. 2005. Automated story extraction from conversational speech. In *Proceedings of the Third International Conference on Knowledge Capture (K-CAP 05)*.

D. Hillard, M. Ostendorf, and E. Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proc. HLT-NAACL*.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *Proceedings of ICASSP-2003*, Hong Kong.

W. Kunz and H. W. J. Ritte. 1970. Issue as elements of information system. Technical Report Working Paper 131, Institute of Urban and Regional Development Research, University of California, Berkeley.

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

J. Niekrasz, M. Purver, J. Dowding, and S. Peters. 2005. Ontology-based discourse understanding for a persistent meeting assistant. In *Proc. of the AAAI Spring Symposium*.

V. Pallotta, J. Niekrasz, and M. Purver. 2005. Collaborative and argumentative models of meeting discussions. In *Proceeding of CMNA-05 workshop on Computational Models of Natural Arguments in IJCAI 05*.

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

R. J. Rienks, D. Heylen, and E. van der Weijden. 2005. Argument diagramming of meeting conversations. In *Multimodal Multiparty Meeting Processing Workshop at the ICMI*.

E. Shriberg and A. Stolcke. 2001. Direct modeling of prosody: An overview of applications in automatic speech processing.

A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf amd T. Schultz, H. Soltau, H. Yu, and K. Zechner. 2001. Advances in automatic meeting record creation and access. In *Proceedings of ICASSP*.

S. Whittaker, R. Laban, and S. Tucker. 2005. Analysing meeting records: An ethnographic study and technological implications. In *Proceedings of MLMI 2005*.

B. Wrede and E. Shriberg. 2003. Spotting hot spots in meetings: Human judgements and prosodic cues. In *Proceedings of EUROSPEECH 2003*.