Modelling and Detecting Decisions in Multi-party Dialogue

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and **Stanley Peters**
Center for the Study of Language and Information
Stanford University
{raquel|frampton|ehlen|mpurver|peters}@stanford.edu

Abstract

We describe a process for automatically detecting decision-making sub-dialogues in transcripts of multi-party, human-human meetings. Extending our previous work on *action item identification*, we propose a structured approach that takes into account the different roles utterances play in the decision-making process. We show that this structured approach outperforms the accuracy achieved by existing decision detection systems based on flat annotations, while enabling the extraction of more fine-grained information that can be used for summarization and reporting.

1 Introduction

In collaborative and organized work environments, people share information and make decisions extensively through multi-party conversations, usually in the form of meetings. When audio or video recordings are made of these meetings, it would be valuable to extract important information, such as the decisions that were made and the trains of reasoning that led to those decisions. Such a capability would allow work groups to keep track of courses of action that were shelved or rejected, and could allow new team members to get quickly up to speed. Thanks to the recent availability of substantial meeting corpora—such as the ISL (Burger et al., 2002), ICSI (Janin et al., 2004), and AMI (McCowan et al., 2005) Meeting Corpora—current research on the structure of decision-making dialogue and its use for automatic decision detection has helped to bring this vision closer to reality (Verbree et al., 2006; Hsueh and Moore, 2007b).

Our aim here is to further that research by applying a simple notion of dialogue structure to the task of automatically detecting decisions in multi-party dialogue. A central hypothesis underlying our approach is that this task is best addressed by taking into account the roles that different utterances play in the decision-making process. Our claim is that this approach facilitates both the detection of regions of discourse where decisions are discussed and adopted, and also the identification of important aspects of the decision discussions themselves, thus opening the way to better and more concise reporting.

In the next section, we describe prior work on related efforts, including our own work on action item detection (Purver et al., 2007). Sections 3 and 4 then present our decision annotation scheme, which distinguishes several types of decision-related dialogue acts (DAs), and the corpus used as data (in this study a section of the AMI Meeting Corpus). Next, in Section 5, we describe our experimental methodology, including the basic conception of our classification approach, the features we used in classification, and our evaluation metrics. Section 6 then presents our results, obtained with a hierarchical classifier that first trains individual *sub-classifiers* to detect the different types of decision DAs, and then uses a *super-classifier* to detect decision regions on the basis of patterns of these DAs, achieving an F-score of 58%. Finally, Section 7 presents some conclusions and directions for future work.

2 Related Work

Recent years have seen an increasing interest in research on decision-making dialogue. To a great extent, this is due to the fact that decisions have

Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pages 156–163,
Columbus, June 2008.

been shown to be a key aspect of meeting speech. User studies (Lisowska et al., 2004; Banerjee et al., 2005) have shown that participants regard decisions as one of the most important outputs of a meeting, while Whittaker et al. (2006) found that the development of an automatic decision detection component is critical to the re-use of meeting archives. Identifying decision-making regions in meeting transcripts can thus be expected to support development of a wide range of applications, such as automatic meeting assistants that process, understand, summarize and report the output of meetings; meeting tracking systems that assist in implementing decisions; and group decision support systems that, for instance, help in constructing group memory (Romano and Nunamaker, 2001; Post et al., 2004; Voss et al., 2007).

Previously researchers have focused on the interactive aspects of argumentative and decision-making dialogue, tackling issues such as the detection of agreement and disagreement and the level of emotional involvement of conversational participants (Hillard et al., 2003; Wrede and Shriberg, 2003; Galley et al., 2004; Gatica-Perez et al., 2005). From a perhaps more formal perspective, Verbree et al. (2006) have created an argumentation scheme intended to support automatic production of argument structure diagrams from decision-oriented meeting transcripts. Only Hsueh and Moore (2007a; 2007b), however, have specifically investigated the automatic detection of decisions.

Using the AMI Meeting Corpus, Hsueh and Moore (2007b) attempt to identify the dialogue acts (DAs) in a meeting transcript that are "decision-related". The authors define these DAs on the basis of two kinds of manually created summaries: an extractive summary of the whole meeting, and an abstractive summary of the decisions made in the meeting. Those DAs in the extractive summary that support any of the decisions in the abstractive summary are then manually tagged as decision-related DAs. They trained a Maximum Entropy classifier to recognize this single DA class, using a variety of lexical, prosodic, dialogue act and topical features. The F-score they achieved was 0.35, which gives a good indication of the difficulty of this task.

In our previous work (Purver et al., 2007), we attempted to detect a particular kind of decision common in meetings, namely *action items*—public commitments to perform a given task. In contrast to the approach adopted by Hsueh and Moore (2007b), we proposed a hierarchical approach where individual classifiers were trained to detect distinct action item-related DA classes (*task description*, *time-frame*, *ownership* and *agreement*) followed by a super-classifier trained on the hypothesized class labels and confidence scores from the individual classifiers that would detect clusters of multiple classes. We showed that this structured approach produced better classification accuracy (around 0.39 F-score on the task of detecting action item regions) than a flat-classifier baseline trained on a single action item DA class (around 0.35 F-score).

In this paper we extend this approach to the more general task of detecting decisions, hypothesizing that—as with action items—the dialogue acts involved in decision-making dialogue form a rather heterogeneous set, whose members co-occur in particular kinds of patterns, and that exploiting this richer structure can facilitate their detection.

3 Decision Dialogue Acts

We are interested in identifying the main conversational units in a decision-making process. We expect that identifying these units will help in detecting regions of dialogue where decisions are made (*decision sub-dialogues*), while also contributing to identification and extraction of specific decision-related bits of information.

Decision-making dialogue can be complex, often involving detailed discussions with complicated argumentative structure (Verbree et al., 2006). Decision sub-dialogues can thus include a great deal of information that is potentially worth extracting. For instance, we may be interested in knowing what a decision is about, what alternative proposals were considered during the decision process, what arguments were given for and against each of them, and last but not least, what the final resolution was.

Extracting these and other potential decision components is a challenging task, which we do not intend to fully address in this paper. This initial study concentrates on three main components we believe constitute the backbone of decision sub-dialogues. A typical decision sub-dialogue consists of three main components that often unfold in sequence. (a)

key	DDA class	description
I	*issue*	utterances introducing the issue or topic under discussion
R	*resolution*	utterances containing the decision that is adopted
RP	– *proposal*	– utterances where the decision adopted is proposed
RR	– *restatement*	– utterances where the decision adopted is confirmed or restated
A	*agreement*	utterances explicitly signalling agreement with the decision made

Table 1: Set of decision dialogue act (DDA) classes

A topic or issue that requires some sort of conclusion is initially raised. (b) One or more proposals are considered. And (c) once some sort of agreement is reached upon a particular resolution, a decision is adopted.

Dialogue act taxonomies often include tags that can be decision-related. For instance, the DAMSL taxonomy (Core and Allen, 1997) includes the tags `agreement` and `commit`, as well as a tag `open-option` for utterances that "suggest a course of action". Similarly, the AMI DA scheme[1] incorporates tags like `suggest`, `elicit-offer-or-suggestion` and `assess`. These tags are however very general and do not capture the distinction between decisions and more general suggestions and commitments.[2] We therefore devised a decision annotation scheme that classifies utterances according to the role they play in the process of formulating and agreeing on a decision. Our scheme distinguishes among three main decision dialogue act (DDA) classes: *issue* (*I*), *resolution* (*R*), and *agreement* (*A*). Class *R* is further subdivided into *resolution proposal* (*RP*) and *resolution restatement* (*RR*). A summary of the classes is given in Table 1.

Annotation of the *issue* class includes any utterances that introduce the topic of the decision discussion. For instance, in example (1) below, the utterances *"Are we going to have a backup?"* and *"But would a backup really be necessary?"* are tagged as *I*. The classes *RP* and *RR* are used to annotate those utterances that specify the resolution adopted—i.e. the decision made. Annotation with the class *RP* includes any utterances where the resolution is initially proposed (like the utterance *"I think maybe we could just go for the kinetic energy. . . "*). Sometimes decision discussions include utterances that sum up the resolution adopted, like the utterance *"Okay, fully kinetic energy"* in (1). This kind of utterance is tagged with the class *RR*. Finally, the *agreement* class includes any utterances in which participants agree with the (proposed) resolution, like the utterances *"Yeah"* and *"Good"* as well as *"Okay"* in dialogue (1).

(1) A: Are we going to have a backup?
Or we do just–
B: But would a backup really be necessary?
A: I think maybe we could just go for the kinetic energy and be bold and innovative.
C: Yeah.
B: I think– yeah.
A: It could even be one of our selling points.
C: Yeah *–laugh–*.
D: Environmentally conscious or something.
A: Yeah.
B: Okay, fully kinetic energy.
D: Good.[3]

Note that an utterance can be assigned to more than one of these classes. For instance, the utterance *"Okay, fully kinetic energy"* is annotated both as *RR* and *A*. Similarly, each decision sub-dialogue may contain more than one utterance corresponding to each class, as we saw above for *issue*. While we do not a priori require each of these classes to be present for a set of utterances to be considered a decision sub-dialogue, all annotated decision sub-dialogues in our corpus include the classes *I*, *RP* and *A*. The annotation process and results are described in detail in the next section.

[1] A full description of the AMI Meeting Corpus DA scheme is available at `http://mmm.idiap.ch/private/ami/annotation/dialogue_acts_manual_1.0.pdf`, after free registration.

[2] Although they can of course be used to aid the identification process—see Section 5.3.

[3] This example was extracted from the AMI dialogue ES2015c and has been modified slightly for presentation purposes.

4 Data: Corpus & Annotation

In this study, we use 17 meetings from the AMI Meeting Corpus (McCowan et al., 2005), a publicly available corpus of multi-party meetings containing both audio recordings and manual transcriptions, as well as a wide range of annotated information including dialogue acts and topic segmentation. Conversations are all in English, but they can include native and non-native English speakers. All meetings in our sub-corpus are driven by an elicitation scenario, wherein four participants play the role of *project manager*, *marketing expert*, *interface designer*, and *industrial designer* in a company's design team. The overall sub-corpus makes up a total of 15,680 utterances/dialogue acts (approximately 920 per meeting). Each meeting lasts around 30 minutes.

Two authors annotated 9 and 10 dialogues each, overlapping on two dialogues. Inter-annotator agreement on these two dialogues was similar to (Purver et al., 2007), with *kappa* values ranging from 0.63 to 0.73 for the four DDA classes. The highest agreement was obtained for class *RP* and the lowest for class *A*.[4]

On average, each meeting contains around 40 DAs tagged with one or more of the DDA subclasses in Table 1. DDAs are thus very sparse, corresponding to only 4.3% of utterances. When we look at the individual DDA sub-classes this is even more pronounced. Utterances tagged as *issue* make up less than 0.9% of utterances in a meeting, while utterances annotated as *resolution* make up around 1.4%—1% corresponding to *RP* and less than 0.4% to *RR* on average. Almost half of DDA utterances (slightly over 2% of all utterances on average) are tagged as belonging to class *agreement*.

We compared our annotations with the annotations of Hsueh and Moore (2007b) for the 17 meetings of our sub-corpus. The overall number of utterances annotated as decision-related is similar in the two studies: 40 *vs.* 30 utterances per meeting on average, respectively. However, the overlap of the annotations is very small leading to negative *kappa* scores. As shown in Figure 1, only 12.22% of utterances tagged with one of our DDA classes correspond to an utterance annotated as decision-related by Hsueh & Moore. While presumably this is a consequence of our different definitions for DDAs, it seems also partially due to the fact that sometimes we disagreed about where decisions were being made. Most of the overlap is found with utterances tagged as *resolution* (*RP* or *RR*). Around 32% of utterances tagged as *resolution* overlap with AMI DDAs, while the overlap with utterances annotated as *issue* and *agreement* is substantially lower—around 7% and 1.5%, respectively. This is perhaps not surprising given their definition of a "decision-related" DA (see Section 2). Classes like *issue* and especially *agreement* shape the interaction patterns of decision-sub-dialogues, but are perhaps unlikely to appear in an extractive summary.[5]

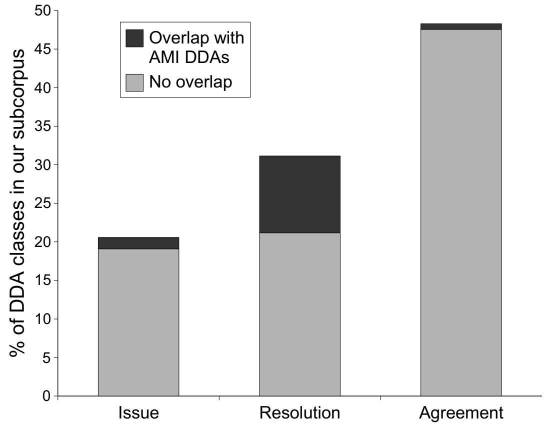

Figure 1: Overlap with AMI annotations

5 Experiments

5.1 Classifiers

Our hierarchical approach to decision detection involves two steps:

1. We first train one independent *sub-classifier* for the identification of each of our DDA classes, using features derived from the properties of the utterances in context (see below).

2. To detect decision sub-dialogues, we then train a *super-classifier*, whose features are the hypothesized class labels and confidence scores

[4] The annotation guidelines we used are available online at `http://godel.stanford.edu/twiki/bin/view/Calo/CaloDecisionDiscussionSchema`

[5] Although, as we shall see in Section 6.2, they contribute to improve the detection of decision sub-dialogues and of other DDA classes.

from the sub-classifiers, over a suitable window.[6]

The super-classifier is then able to "correct" the DDA classes hypothesized by the sub-classifiers on the basis of richer contextual information: if a DA is classified as positive by a sub-classifier, but negative by the super-classifier, then this sub-classification is "corrected", i.e. it is changed to negative. Hence this hierarchical approach takes advantage of the fact that within decision sub-dialogues, our DDAs can be expected to co-occur in particular types of patterns.

We use the linear-kernel support vector machine classifier SVMlight (Joachims, 1999) in all classification experiments.

5.2 Evaluation

In all cases we perform 17-fold cross-validation, each fold training on 16 meetings and testing on the remaining one.

We can evaluate the performance of our approach at three levels: the accuracy of the sub-classifiers in detecting each of the DDA classes, the accuracy obtained in detecting DDA classes after the output of the sub-classifiers has been corrected by the super-classifier, and the accuracy of the super-classifier in detecting decision sub-dialogues. For the DDA identification task (both uncorrected and corrected) we use the same lenient-match metric as Hsueh and Moore (2007b), which allows a margin of 20 seconds preceding and following a hypothesized DDA.[7] We take as reference the results they obtained on detecting their decision-related DAs.

For the evaluation of the decision sub-dialogue detection task, we follow (Purver et al., 2007) and use a windowed metric that divides the dialogue into 30-second windows and evaluates on a per window basis. As a baseline for this task, we compare the performance of our hierarchical approach to a flat classification approach, first using the flat annotations of Hsueh and Moore (2007a) that only include a single DDA class, and second using our annotations, but for the binary classification of whether an utterance is decision-related or not, without distinguishing among our DDA sub-classes.

[6]The width of this window is estimated from the training data and corresponds to the average length in utterances of a decision sub-dialogue—25 in our sub-corpus.

[7]Note that here we only give credit for hypotheses based on a 1–1 mapping with the gold-standard labels.

5.3 Features

To train the DDA sub-classifiers we extracted utterance features similar to those used by Purver et al. (2007) and Hsueh and Moore (2007b): lexical unigrams and durational and locational features from the transcripts; prosodic features extracted from the audio files using Praat (Boersma, 2001); general DA tags and speaker information from the AMI annotations; and contextual features consisting of the same set of features from immediately preceding and following utterances. Table 2 shows the full feature set.

Lexical	unigrams after text normalization
Utterance	length in words, duration in seconds, percentage of meeting
Prosodic	pitch & intensity min/max/mean/dev, pitch slope, num of voice frames
DA	AMI dialogue act class
Speaker	speaker id & AMI speaker role
Context	features as above for utterances u +/- 1… u +/- 5

Table 2: Features for decision DA detection

6 Results

6.1 Baseline

On the task of detecting decision-related DAs, Hsueh and Moore (2007b) report an F-score of 0.33 when only lexical features are employed. Using a combination of different features allows them to boost the score to 0.35. Although the differences both in definition and prior distribution between their DAs and our DDA classes make direct comparisons unstraightforward (see Sec. 4), we consider this result a baseline for the DDA detection task.

As a baseline system for the decision sub-dialogue detection task, we use a flat classifier trained on the word unigrams of the current utterance (lexical features) and the unigrams of the immediately preceding and following utterances (+/-1-utterance context). Table 3 shows the accuracy per 30-second window obtained when a flat classifier is applied to AMI annotations and to our own annotations, respectively.[8] In general, the flat classifiers yield high recall (over 90%) but rather low precision (below 35%).

[8]Note that the task of detecting decision sub-dialogues is not directly addressed by (Hsueh and Moore, 2007b).

As can be seen, using our DA annotations (CALO DDAs) with all sub-classes merged into a single class yields better results than using the AMI DDA flat annotations. The reasons behind this result are not entirely obvious. In principle, our annotated DDAs are by definition less homogeneous than the AMI DDAs, which could lead to a lower performance in a simple binary approach. It seems however that the regions that contain our DDAs are easier to detect than the regions that contain AMI DDAs.

Flat classifier	Re	Pr	F1
AMI DDAs	.97	.21	.34
CALO DDAs	.96	.34	.50

Table 3: Flat classifiers with lexical features and +/–1-utterance context

6.2 Hierarchical Results

Performance of the hierarchical classifier with lexical features and +/- 1-utterance context is shown in Table 4. The results of the super-classifier can be compared directly to the baseline flat classifier of Table 3. We can see that the use of the super-classifier to detect decision sub-dialogues gives a significantly improved performance over the flat approach. This is despite low sub-classifier performance, especially for the classes with very low frequency of occurrence like *RR*. Precision for decision sub-dialogue detection improves around 0.5 points ($p < 0.05$ on an paired t-test), boosting F-scores to 0.55 ($p < 0.05$). The drop in recall from 0.96 to 0.91 is not statistically significant.

	sub-classifiers				super
	I	RP	RR	A	classifier
Re	.25	.44	.09	.88	.91
Pr	.21	.24	.14	.18	.39
F1	.23	.31	.11	.30	.55

Table 4: Hierarchical classifier with lexical features and +/–1-utterance context

We investigated whether we could improve results further by using additional features, and found that we could. The best results obtained with the hierarchical classifier are shown in Table 5. We applied feature selection to the features shown in Table 2 using *information gain* and carried out several trial classifier experiments. Like Purver et al. (2007) and (Hsueh and Moore, 2007b), we found that lexical features increase classifier performance the most.

As DA features, we used the AMI DA tags `elicit-assessment`, `suggest` and `assess` for classes *I* and *A*; and tags `suggest`, `fragment` and `stall`, for classes *RP* and *RR*. Only the DA features for the *Resolution* sub-classes (*RP* and *RR*) gave significant improvements ($p < 0.05$). Utterance and speaker features were found to improve the recall of the sub-classes significantly ($p < 0.05$), and the precision of the super-classifier ($p < 0.05$). As for prosodic information, we found minimum and maximum intensity to be the most generally predictive, but although these features increased recall, they caused precision and F-scores to decrease.

When we experimented with contextual features (i.e. features from utterances before and after the current dialogue act), we only found lexical contextual features to be useful. With the current dataset, for classes *I*, *RP* and *RR*, the optimal amount of lexical contextual information turned out to be +/- 1 utterances, while for class *A* increasing the amount of lexical contextual information to +/-5 utterances yielded better results, boosting both precision and F-score ($p < 0.05$). Speaker, utterance, DA and prosodic contextual features gave no improvement.

The scores on the left hand side of Table 5 show the best results obtained with the sub-classifiers for each of the DDA classes. We found however that the super-classifier was able to improve over these results by correcting the hypothesized labels on the basis of the DDA patterns observed in context (see the corrected results on Table 5). In particular, precision increased from 0.18 to 0.20 for class *I* and from 0.28 to 0.31 for class *RP* (both results are statistically significant, $p < 0.05$). Our best F-score for class *RP* (which is the class with the highest overlap with AMI DDAs) is a few points higher than the one reported in (Hsueh and Moore, 2007b)—0.38 *vs*. 0.35, respectively.

Next we investigated the contribution of the class *agreement*. Although this class is not as informative for summarization and reporting as the other DDA classes, it plays a key role in the interactive process that shapes decision sub-dialogues. Indeed, including this class helps to detect other more contentful DDA classes such as *issue* and *resolution*.

	sub-classifiers				corr. sub-classifiers				corr. sub. w/o A			super	super
	I	RP	RR	A	I	RP	RR	A	I	RP	RR	w/o A	with A
Re	.45	.49	.18	.55	.43	.48	.18	.55	.43	.48	.18	.91	.88
Pr	.18	.28	.14	.30	.20	.31	.14	.30	.18	.30	.14	.36	.43
F1	.25	.36	.16	.39	.28	.38	.16	.39	. 26	.37	.16	.52	.58

Table 5: Hierarchical classifier with uncorrected and corrected results for sub-classifiers, with and w/o class A; lexical, utterance, and speaker features; +/–1-utt lexical context for I-RP-RR and +/–5-utt lexical context for A.

Table 5 also shows the results obtained with the hierarchical classifier when class *A* is ignored. In this case the small correction observed in the precision of classes *I* and *RP* w.r.t. the original output of the sub-classifiers is not statistically significant. The performance of the super-classifier (sub-dialogue detection) also decreases significantly in this condition: 0.43 *vs*. 0.36 precision and 0.58 *vs*. 0.52 F-score ($p < 0.05$).

6.3 Robustness to ASR output

Finally, since the end goal is a system that can automatically extract decisions from raw audio and video recordings of meetings, we also investigated the impact of ASR output on our approach. We used SRI's Decipher (Stolcke et al., 2008)[9] to produce word confusion networks for our 17 meeting sub-corpus and then ran our detectors on the WCNs' best path. Table 6 shows a comparison of F-scores. The two scores shown for the super-classifier correspond to using the best feature set *vs*. using only lexical features. When ASR output is used, the results for the DDA classes decrease between 6 and 11 points. However, the performance of the super-classifier does not experience a significant degradation (the drop in F-score from 0.58 to 0.51 is not statistically significant). The results obtained with the hierarchical detector are still significantly higher than those achieved by the flat classifier (0.51 *vs*. 0.50, $p < 0.05$).

F1	I	RP	RR	A	super	flat
WCNs	.22	.30	.08	.28	.51/.51	.50
Manual	.28	.38	.16	.39	.58/.55	.50

Table 6: Comparison of F-scores obtained with WCNs and manual transcriptions

[9] Stolcke et al. (2008) report a word error rate of 26.9% on AMI meetings.

7 Conclusions & Future Work

We have shown that our earlier approach to action item detection can be successfully applied to the more general task of detecting decisions. Although this is indeed a hard problem, we have shown that results for automatic decision-detection in multi-party dialogue can be improved by taking account of dialogue structure and applying a hierarchical approach. Our approach consists in distinguishing between the different roles utterances play in the decision-making process and uses a hierarchical classification strategy: individual sub-classifiers are first trained to detect each of the DDA classes; then a super-classifier is used to detect patterns of these classes and identify decisions sub-dialogues. As we have seen, this structured approach outperforms the accuracy achieved by systems based on flat classifications. For the task of detecting decision sub-dialogues we achieved 0.58 F-score in initial experiments—a performance that proved to be rather robust to ASR output. Results for the individual sub-classes are still low and there is indeed a lot of room for improvement. In future work, we plan to increase the size of our data-set, and possibly extend our set of DDA classes, by for instance including a *disagreement* class, in order to capture additional properties of the decision-making process.

We believe that our structured approach can help in constructing more concise and targeted reports of decision sub-dialogues. An immediate further extension of the current work will therefore be to investigate the automatic production of useful descriptive summaries of decisions.

Acknowledgements We are thankful to the three anonymous SIGdial reviewers for their helpful comments and suggestions. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-07-D-0185/0004. Any opinions, find-

ings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Satanjeev Banerjee, Carolyn Rosé, and Alex 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 10th International Conference on Human-Computer Interaction*.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10).

Susanne Burger, Victoria MacLaren, and Hua Yu. 2002. The ISL Meeting Corpus: The impact of meeting type on speech style. In *Proceedings of the 7th International Conference on Spoken Language Processing (INTERSPEECH - ICSLP)*, Denver, Colorado.

Mark Core and James Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In D. Traum, editor, *Proceedings of the 1997 AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth 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 Association for Computational Linguistics (ACL)*.

Daniel Gatica-Perez, Ian McCowan, Dong Zhang, and Samy Bengio. 2005. Detecting group interest level in meetings. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Dustin Hillard, Mari Ostendorf, and Elisabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, Edmonton, Alberta, May.

Pei-Yun Hsueh and Johanna Moore. 2007a. What decisions have you made?: Automatic decision detection in meeting conversations. In *Proceedings of NAACL/HLT*, Rochester, New York.

Pey-Yun Hsueh and Johanna Moore. 2007b. Automatic decision detection in meeting speech. In *Proceedings of MLMI 2007*, Lecture Notes in Computer Science. Springer-Verlag.

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Marcías-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. 2004. The ICSI meeting project: Resources and research. In *Proceedings of the 2004 ICASSP NIST Meeting Recognition Workshop*.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*. MIT Press.

Agnes Lisowska, Andrei Popescu-Belis, and Susan Armstrong. 2004. User query analysis for the specification and evaluation of a dialogue processing and retrieval system. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Iain McCowan, Jean Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI Meeting Corpus. In *Proceedings of Measuring Behavior 2005, the 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, Netherlands.

Wilfried M. Post, Anita H.M. Cremers, and Olivier Blanson Henkemans. 2004. A research environment for meeting behaviour. In *Proceedings of the 3^{rd} Workshop on Social Intelligence Design*.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

Nicholas C. Romano, Jr. and Jay F. Nunamaker, Jr. 2001. Meeting analysis: Findings from research and practice. In *Proceedings of the 34th Hawaii International Conference on System Sciences*.

Andreas Stolcke, Xavier Anguera, Kofi Boakye, Özgür Çetin, Adam Janin, Matthew Magimai-Doss, Chuck Wooters, and Jing Zheng. 2008. The icsi-sri spring 2007 meeting and lecture recognition system. In *Proceedings of CLEAR 2007 and RT2007*. Springer Lecture Notes on Computer Science.

Daan Verbree, Rutger Rienks, and Dirk Heylen. 2006. First steps towards the automatic construction of argument-diagrams from real discussions. In *Proceedings of the 1st International Conference on Computational Models of Argument*, volume 144, pages 183–194. IOS press.

Lynn Voss, Patrick Ehlen, and the DARPA CALO MA Project Team. 2007. The CALO Meeting Assistant. In *Proceedings of NAACL-HLT*, Rochester, NY, USA.

Steve Whittaker, Rachel Laban, and Simon Tucker. 2006. Analysing meeting records: An ethnographic study and technological implications. In *MLMI 2005, Revised Selected Papers*.

Britta Wrede and Elizabeth Shriberg. 2003. Spotting "hot spots" in meetings: Human judgements and prosodic cues. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, Geneva, Switzerland.