Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity

Trung H. Bui[1], **Matthew Frampton**[1], **John Dowding**[2], and **Stanley Peters**[1]

[1]Center for the Study of Language and Information, Stanford University
{thbui|frampton|peters}@stanford.edu

[2]University of California/Santa Cruz
jdowding@ucsc.edu

Abstract

We use directed graphical models (DGMs) to automatically detect decision discussions in multi-party dialogue. Our approach distinguishes between different dialogue act (DA) types based on their role in the formulation of a decision. DGMs enable us to model dependencies, including sequential ones. We summarize decisions by extracting suitable phrases from DAs that concern the issue under discussion and its resolution. Here we use a semantic-similarity metric to improve results on both manual and ASR transcripts.

1 Introduction

In work environments, people share information and make decisions in multi-party conversations known as meetings. The demand for systems that can automatically process, understand and summarize information contained in audio and video recordings of meetings is growing rapidly. Our own research, and that of other contemporary projects (Janin et al., 2004), aim at meeting this demand.

At present, we are focusing on the automatic detection and summarization of decision discussions. Our approach for detecting decision discussions involves distinguishing between different dialogue act (DA) types based on their role in the decision-making process. Two of these types are DAs which describe the *Issue* under discussion, and DAs which describe its *Resolution*. To summarize a decision discussion, we identify words and phrases in the Issue and Resolution DAs, which can be used to produce a concise, descriptive summary.

This paper describes new experiments in both detecting and summarizing decision discussions. In the detection stage, we investigate the use of Directed Graphical Models (DGMs). DGMs are attractive because they can be used to model sequence and dependencies between predictor variables. In the summarization stage, we attempt to improve phrase selection with a new feature that measures the level of semantic similarity between candidate Issue phrases and Resolution utterances, and vice-versa. The feature is generated by a semantic-similarity metric which uses WordNet as a knowledge source. The motivation is that ordinarily, the Issue and Resolution components in a decision summary should be semantically similar.

The paper proceeds as follows. Firstly, Section 2 describes related work, and Section 3, our data-set and annotation scheme for decision discussions. Section 4 then reports our decision detection experiments using DGMs, and Section 5, the summarization experiments. Finally, Section 6 draws conclusions and proposes ideas for future work.

2 Related Work

User studies (Banerjee et al., 2005) have confirmed that meeting participants consider decisions to be one of the most important meeting outputs, and (Whittaker et al., 2006) found that the development of an automatic decision detection component is critical to the re-use of meeting archives. With the new availability of substantial meeting corpora such as the AMI corpus (McCowan et al., 2005), recent years have therefore seen an increasing amount of research on decision-making dialog. This research has tackled issues such as the automatic detection of agreement and disagreement (Galley et al., 2004), and of the

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 235–243,
Queen Mary University of London, September 2009. ©2009 Association for Computational Linguistics

level of involvement of conversational participants (Gatica-Perez et al., 2005). In addition, (Verbree et al., 2006) created an argumentation scheme intended to support automatic production of argument structure diagrams from decision-oriented meeting transcripts. As yet, there has been relatively little work which specifically addresses the automatic detection and summarization of decisions.

Decision discussion detection: (Hsueh and Moore, 2007) used the AMI Meeting Corpus, and attempted to automatically identify DAs in meeting transcripts which are "decision-related". For each meeting, two manually created summaries were used to judge which DAs were decision-related: an extractive summary of the whole meeting, and an abstractive summary of its decisions. Those DAs in the extractive summary which support any of the decisions in the abstractive summary were manually tagged as decision-related. (Hsueh and Moore, 2007) then trained a Maximum Entropy classifier to recognize this single DA class, using a variety of lexical, prosodic, DA and conversational topic features. They achieved an F-score of 0.35.

Unlike (Hsueh and Moore, 2007), (Fernández et al., 2008b) made an attempt at modelling the structure of decision-making dialogue. The authors designed an annotation scheme that takes account of the different roles which utterances can play in the decision-making process—for example it distinguishes between DDAs (decision DAs) which initiate a discussion by raising an issue, those which propose a resolution, and those which express agreement for a proposed resolution. The authors annotated a portion of the AMI corpus, and then applied what they refer to as "hierarchical classification". Here, one *sub-classifier* per DDA class hypothesizes occurrences of that DDA class, and then based on these hypotheses, a *super-classifier* determines which regions of dialogue are decision discussions. All of the classifiers, (sub and super), were linear kernel binary Support Vector Machines (SVMs). Results were better than those obtained with (Hsueh and Moore, 2007)'s approach—the F1-score for detecting decision discussions in manual transcripts was .58 *vs.* .50. Note that (Purver et al., 2007) had previously pursued the same basic approach as (Fernández et al., 2008b) in order to detect action items.

In this paper, we build on the promising results of (Fernández et al., 2008b), by using Directed Graphical Models (DGMs) in place of SVMs. DGMs are attractive because they provide a natural framework for modelling sequence and dependencies between variables including the DDAs. We are especially interested in whether DGMs better exploit non-lexical features. (Fernández et al., 2008b) obtained much more value from lexical than non-lexical features (and indeed no value at all from prosodic features), but lexical features have disadvantages. In particular, they can be domain specific, increase the size of the feature space dramatically, and deteriorate more than other features in quality when ASR is poor.

Decision summarization: Recent years have seen research on spoken dialogue summarization (e.g. (Zechner, 2002)). Most has attempted to generate summaries of full dialogues, but some very recent research has focused on specific dialogue events, namely action items (Purver et al., 2007), and decisions (Fernández et al., 2008a).

(Fernández et al., 2008a) used the DDA annotation scheme mentioned above, and began by extracting the DDAs which raise issues or provide accepted resolutions. Only manual transcripts were used and the DDAs were extracted by hand rather than automatically. The next step was to parse each DDA with a general rule-based parser (Dowding et al., 1993), producing multiple short fragments rather than one full utterance parse. Then, for each DDA, an SVM regression model used various features (including parse, semantic and lexical features) to select the fragment which was most likely to appear in a gold-standard extractive decision summary. The entire manual utterance transcriptions were used as the baseline, and although the SVM's precision was high, it was not enough to offset the baseline's perfect recall, and so its F-score was lower. The "Oracle", which always chooses the fragment with the highest F1-score produced very good results. This motivates deeper investigation into how to improve the fragment/parse selection phase, and so we assess the usefulness of a semantic-similarity feature for the SVM. We conduct experiments with ASR as well as manual transcripts.

3 Data

For the experiments reported in this study, we used 17 meetings from the AMI Meeting Corpus (McCowan et al., 2005), a freely available corpus of multi-party meetings with both audio and video recordings, and a wide range of annotated information including DAs and topic segmentation. Conversations are in English, but some participants are non-native English speakers. The meetings last around 30 minutes each, and are scenario-driven, wherein four participants play different roles in a company's design team: *project manager*, *marketing expert*, *interface designer* and *industrial designer*.

3.1 Modelling Decision Discussions

We use the same annotation scheme as (Fernández et al., 2008b) to model decision-making dialogue. As stated in Section 2, this scheme distinguishes between a small number of DA types based on the role which they perform in the formulation of a decision. Apart from improving the initial detection of decision discussions (Fernández et al., 2008b), such a scheme also aids their subsequent summarization, because it indicates which utterances contain particular types of information.

The annotation scheme is based on the observation that a decision discussion contains the following main structural components: (a) a topic or issue requiring resolution is raised, (b) one or more possible resolutions are considered, (c) a particular resolution is agreed upon and so becomes the decision. Hence the scheme distinguishes between 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*). *I* utterances introduce the topic of the decision discussion, examples being *"Are we going to have a backup?"* and *"But would a backup really be necessary?"* in Dialogue 1. On the other hand, *R* utterances specify the resolution which is ultimately adopted as the decision. *RP* utterances propose this resolution (e.g. *"I think maybe we could just go for the kinetic energy..."*), while *RR* utterances close the discussion by confirming/summarizing the decision (e.g. *"Okay, fully kinetic energy"*) . Finally, *A* utterances agree with the proposed resolution, signalling that it is adopted as the decision, (e.g. *"Yeah"*, *"Good"* and *"Okay"*). Note that an utterance can be assigned to more than one DDA class, and within a decision discussion, more than one utterance can be assigned to the same DDA class.

We use both manual and ASR one-best transcripts[1] in the experiments described here. DDA annotations were first made on the manual transcripts, and then transferred onto the ASR transcripts. Inter-annotator agreement was satisfactory, with kappa values ranging from .63 to .73 for the four DDA classes. Due to different segmentation, the manual and ASR transcripts contain a total of 15,680 and 8,357 utterances respectively, and on average, 40 and 33 DDAs per meeting. Hence DDAs are slightly less sparse in the ASR transcripts: for all DDAs, 6.7% *vs.* 4.3% of the total number of utterances, for *I*, 1.6% *vs.* 0.9%, for *RP*, 2% *vs.* 1%, for *RR*, 0.5% *vs.* 0.4%, and for *A*, 2.6% *vs.* 2%.

(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.[2]

4 Decision Discussion Detection using Directed Graphical Models

A directed graphical model (DGM) *M*, (see Murphy (2002)), is a directed acyclic graph consisting of nodes which represent random variables, arcs which represent dependencies among these variables, and a probability distribution P over the variables. Let $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ be a set of random variables that are associated with nodes in a DGM and $Pa(X_i)$ be parents of X_i. The probability distribution of the model M satisfies:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n}(P(X_i)|Pa(X_i))$$

When a DGM is used as a classifier, the goal is to correctly infer the value of the class node $X_c \in \mathbf{X}$ given a vector of values for the observed node(s)

[1] We used SRI's Decipher for which (Stolcke et al., 2008) reports a word error rate of 26.9% on AMI meetings.

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

$X_o \subseteq \mathbf{X} \setminus X_c$. This is done by using M to find the value of X_c which gives the highest conditional probability $P(X_c|X_o)$.

To detect each individual DDA class, we examined the four simple DGMs in Figure 1 (see Appendix). The DDA node is binary where value 1 indicates the presence of a DDA and 0 its absence. The evidence node (E) is a multi-dimensional vector of observed values of non-lexical features. These include utterance features (UTT) such as length in words, duration in milliseconds, position within the meeting (as percentage of elapsed time), manually annotated dialogue act (DA) features[3] such as *inform*, *assess*, *suggest*, and prosodic features (PROS) such as energy and pitch. These features are the same as the non-lexical features used by Fernández et al. (2008b). The hidden component node (C) represents the distribution of observable evidence E as a single Gaussian in the *-sim* models, and a mixture in the *-mix* models. For the -mix models, the number of Gaussian components is hand-tuned during the training phase.

More complex models are constructed from the four simple models in Figure 1 to allow for dependencies between different DDAs. For example, the model in Figure 2 (see Appendix) generalizes Figure 1c with arcs connecting the DDA classes based on analysis of the annotated AMI data.

4.1 Experiments

The DGM classifiers in Figures 1 and 2 were implemented in Matlab using the BNT software[4]. Since the current BNT version does not support multiple time series training for fully observable Dynamic Bayesian Networks (DBNs), we extended the software for training models using this structure (e.g., Figure 1c and Figure 2).

A DGM classifier is considered to have hypothesized a DDA if the marginal probability of its DDA node is above a hand-tuned threshold. We tested the DGMs on manual and ASR transcripts in a 17-fold cross-validation, and evaluated their performance on both a per-utterance basis, and also with the same lenient-match metric as Fernández et al. (2008b). This allows a margin of 20 seconds preceding and following a hypothesized DDA, and so we refer to it as the 40 second metric. In addition, we hypothesized decision discussion regions using the DGM output and the following two simple rules:

- A decision discussion region begins with an *Issue* DDA.
- A decision discussion region contains at least one *Issue* DDA and one *Resolution* DDA.

To evaluate the accuracy of these hypothesized regions, like Fernández et al. (2008b), we divided the dialogue into 30-second windows and evaluated on a per window basis.

4.2 Results

Tables 1 and 2 show the F1-scores for each DGM when using the best feature sets (I: UTT+DA+PROS, RP: UTT+DA, RR: UTT, A: UTT+DA). The BN-mix model gives the highest F1-score for *A* on both evaluation metrics, and the DBN-mix model, the highest for *I*, *RP*, and *RR*, but there are no statistically significant differences between any of the alternative DGMs.

Classifier	I	RP	RR	A
BN-mix	.09	.09	.04	.19
DBN-mix	.16	.14	.05	.17
BN-sim	.12	.09	.04	.17
DBN-sim	.15	.11	.04	.16

Table 1: F1-score (per utterance) of the DGMs using the best combination of non-lexical features.

Classifier	I	RP	RR	A
BN-mix	.19	.24	.07	.38
DBN-mix	.27	.24	.07	.32
BN-sim	.23	.22	.06	.36
DBN-sim	.25	.22	.06	.31

Table 2: F1-score (40 seconds) of the DGMs using the best combination of non-lexical features.

To determine whether modeling dependencies between DDAs improves performance, we experimented with the DGMs that are generalized from the DBN-sim (Figure 2) and DBN-mix models. The F1-scores did not improve for *I*, *RP*, and *RR*, while for *A*, the DGM generalized from DBN-sim gave a .03 improvement according to the 40 seconds metric, but this was not statistically significant.

For each DDA, Table 3 compares the results of the best DGM and the hierarchical SVM classification method of Fernández et al. (2008b) (see

[3] We use the AMI DA annotations. These are only available for manual transcripts.

[4] http://www.cs.ubc.ca/∼murphyk/Software/BNT/bnt.html

Section 2). The DGM performs better for all DDAs on both evaluation metrics ($p < 0.005$). Note that while prosodic features proved useless to SVM classifiers (Fernández et al. (2008b)), with DGMs, they have some predictive power.

		Per utterance			40 seconds		
Classifier	DDA	Pr	Re	F1	Pr	Re	F1
SVM	I	.03	.62	.05	.04	.89	.08
DGM		.11	.28	.16	.20	.44	.27
SVM	RP	.03	.60	.07	.05	.90	.10
DGM		.09	.35	.14	.16	.57	.24
SVM	RR	.01	.49	.02	.01	.80	.03
DGM		.02	.42	.05	.04	.58	.07
SVM	A	.05	.70	.10	.07	.90	.13
DGM		.13	.31	.19	.29	.55	.38

Table 3: Performance of the DGM classifier vs. the SVM classifier. Both use the best combination of non-lexical features.

We also generated results without DA features. Here, the best F1-scores for *I*, *RP*, and *A* degrade between .07 and .09 ($p < 0.05$), but they are still higher than the equivalent SVM results with DA features. Since (Fernández et al., 2008b) report that lexical features are the most useful for the SVM classifiers, it will be interesting to see how well the DGMs perform when they use lexical as well as non-lexical features.

Detecting DDAs in ASR transcripts: Table 4 compares the DGM F1-scores when using ASR one-best and manual transcripts. The DGMs perform well on ASR output. For *I* and *RP*, the results on ASR are actually higher, perhaps because the DDAs are less sparse. In the absence of DA features, prosodic features improve the performance for *A* in both sources.

	UTT				UTT+PROS			
	I	RP	RR	A	I	RP	RR	A
ASR	.20	.21	.06	.24	.16	.24	.07	.28
Man	.18	.17	.07	.27	.16	.15	.05	.30

Table 4: F1-scores (40 seconds) computed using ASR one-best vs. manual transcriptions.

Detecting decision discussion regions: Table 5 shows that according to the 30-second window metric, rule-based classification with DGM output compares well with hierarchical SVM classification (Fernández et al., 2008b). In fact, even when the latter uses lexical as well as non-lexical features, its F1-score is still about the same as the DGM-based classifier. Our future work will involve dispensing with the rule-based approach and designing a DGM which can detect decision discussion regions.

Classifier	Pr	Re	F1
SVM	.35	.88	.50
DGM	.39	.93	.55

Table 5: Results in detecting decision discussion regions for the SVM super-classifier and rule-based DGM classifier, both using the best combination of non-lexical features.

5 Decision Summarization

We now turn to the task of extracting useful phrases for summarization. Since a summary of a decision discussion should minimally contain the issue under discussion, and its resolution, we leave *Agreement (A)* utterances aside, and concentrate on extracting phrases from *Issues (I)* and *Resolutions (R)*.

Our basic approach is the same taken in (Fernández et al., 2008a): The WCN[5] of each *I* and *R* utterance is parsed by the Gemini parser (Dowding et al., 1993) to produce multiple short fragments, and then an SVM regression model uses certain features in order to select the parse that is most likely to match a gold-standard extractive summary. Our work is new in two respects: summarizing from ASR output in addition to manual transcriptions, and using a semantic-similarity feature in the SVM. This new feature is generated using Ted Pedersen's semantic-similarity package (Pedersen, 2002), and is motivated by the fact that ordinarily the *Issue* summary should be semantically similar to the *Resolution* and vice versa.

The next section describes the lexical resources used by Gemini, and Section 5.2, the metric for calculating semantic similarity.

5.1 Open-Domain Semantic Parser

Since human-human spoken dialogue, especially after being processed by an imperfect recognizer, is likely to be highly ungrammatical, we have developed a semantic parser that only attempts to find basic predicate-argument structures of the major phrase types (S, VP, NP, and PP) and has access to a broad-coverage lexicon. To build a broad-coverage lexicon, we used publicly available lexical resources for English, including COMLEX,

[5] When using manual transcripts, we create "dummy WCNs": WCNs with a single path.

VerbNet, WordNet, and NOMLEX.

COMLEX provides detailed syntactic information for the 40k most common words of English, and VerbNet, detailed semantic information for verbs, including verb class, verb frames, thematic roles, mappings of syntactic position to thematic roles, and selection restrictions on thematic role fillers. From WordNet we extracted another 15K nouns and the semantic class information for all nouns. These semantic classes were hand-aligned to the selectional classes used in VerbNet, based on the upper ontology of EuroWordNet. NOMLEX provides syntactic information for event nominalizations, and information for mapping the noun arguments to the corresponding verb syntactic positions.

These resources were combined and converted to the Prolog-based format used in the Gemini framework, which includes a fast bottom-up robust parser in which syntactic and semantic information is applied interleaved. Gemini can compute parse probabilities on the context-free skeleton of the grammar. In the experiments described here these parse probabilities are trained on Switchboard tree-bank data.

5.2 Semantic Similarity Metric: Normalized Path Length

Ted Pedersen's semantic similarity package (Pedersen, 2002) can be used to apply a number of different metrics that use WordNet as a knowledge base. The metric used here, *Normalized Path Length* (Leacock and Chodorow, 1998), defines the semantic similarity *sim* between words w_1 and w_2 as:

$$sim_{c_1,c_2} = -\log \frac{len(c_1, c_2)}{2 \times D} \quad (1)$$

where c_1 and c_2 are concepts corresponding to w_1 and w_2, $len(c_1, c_2)$ is the length of the shortest path between them, and D is the maximum depth of the taxonomy.

5.3 Experiments

Data: For the manual transcripts in our subcorpus, the average length in words of *I* and *R* utterances is 12.2 and 11.9 respectively, and for the ASR, 22.4 and 18.1. To provide a gold-standard, phrases from *I* and *R* utterances in the manual transcriptions were annotated as summary-worthy. The aim was to select those phrases which should appear in an extractive summary, or could be the basis of a generated abstractive summary. As a general guideline, we tried to select the phrase(s) which describe the issue/resolution as succinctly as possible. This does not include phrases which express the speaker's attitude towards the issue/resolution. Dialogue 2 is an example where square brackets indicate which phrases were selected as summary-worthy.

(2) A:(*I*) So we we're looking at [*sliders for both volume and channel change*]
B:(*R*) I was thinking kind of [*just for the volume*]

Regression models: We use *SVMlight* (Joachims, 1999) to learn separate SVM regression models for *Issues* and *Resolutions*. These rank the Gemini parses for each utterance according to their likelihood of matching the gold-standard summary. The top-ranked parse is then entered into the automatically-generated decision summary.

Features: We train the regression models with various types of feature (see Table 6), including properties of the WCN paths, parse, semantic and lexical features. As lexical features are likely to be more domain-specific, and they dramatically increase size of the feature space, we prefer to avoid them if possible.

To generate the semantic-similarity feature for an I/R parse, we compute its semantic similarity with the full transcripts of each of the R/I utterances within the same decision discussion. The feature's value is then equal to the greatest of the resulting semantic-similarity scores. Since Ted Pedersen's package operates on the noun portion of WordNet, we must first extract all of the nouns in the parse/utterance transcription. Next, we form all of the possible pairs containing one noun from the parse, and one from the utterance transcription. Then we compute the semantic similarity for each pair, and take their sum to be the level of semantic similarity between the parse and the utterance transcription. We experimented with averaging rather than summing these scores, but the resulting semantic-similarity feature was less predictive.

Evaluation: The models are evaluated in 10-fold cross-validations using the same metric as (Fernández et al., 2008a): Recall is the total proportion of the gold-standard extractive summary

WCN	phrase length (WCN arcs) start/end point (absolute & percentage)
Parse	parse probability phrase type (S/VP/NP/PP)
Semantic	main verb VerbNet class head noun WordNet synset
Sem-sim	Normalized Path Length
Lexical	main verb, head noun

Table 6: Features for parse fragment ranking

	Issue			*Resolution*		
	Re	Pr	F1	Re	Pr	F1
Baseline	1.0	.50	.67	1.0	.60	.75
Oracle	.77	.96	.85	.74	.99	.84
WCN,parse,sem	.63	.69	.66	.61	.66	.64
+ sem-sim	.65	.71	.68	.64	.69	.67
+ lexical	.65	.67	.66	.65	.70	.67

Table 7: Parse ranking results for *I* & *R* Utterances using manual transcriptions.

covered by the selected parse; precision is the total proportion of the chosen parse which overlaps with the gold-standard summary. The baseline is the entire transcription, and we also compare to an "oracle" that always chooses a parse with the highest F1-score. Note that we use the extractive summaries from the manual transcriptions as the gold-standard for the evaluation of the results obtained with ASR.

Results and analysis: Results with manual transcriptions are shown in Table 7, and those with ASR, in Table 8. In all cases, when starting with a feature set containing WCN, parse and semantic features, the F1-score is improved by adding the semantic-similarity feature. For *Issues*, the F1-score improves from .66 to .68 with manual transcripts, and from .30 to .32 with ASR. The improvements for *Resolutions* are highly significant: with manual transcripts, the F1 score increases from .64 to .67 ($p < 0.005$), and with ASR, from .33 to .37 ($p < 0.005$). Note that the further addition of lexical features only produces a significant improvement in the case of *I* summarization with ASR.

Compared to the full transcript baseline, we achieve higher F1-scores for *Issues*—.68 *vs.* .67 with manual transcriptions, and .35 *vs.* .31 with ASR—but slightly lower for *Resolutions*. There remains a fairly large gap between our best scores and their corresponding oracles (especially with ASR), and so there may still be potential for substantial improvement.

	Issue			*Resolution*		
	Re	Pr	F1	Re	Pr	F1
Baseline	.77	.20	.31	.80	.27	.40
Oracle	.61	.87	.72	.59	.91	.72
WCN,parse,sem	.28	.33	.30	.31	.35	.33
+ sem-sim	.30	.34	.32	.35	.38	.36
+ lexical	.35	.35	.35	.34	.39	.37

Table 8: Parse ranking results for *I* & *R* Utterances using ASR.

6 Conclusions and Future Work

This paper has presented work on the detection and summarization of decision discussions in multi-party dialogue. In the detection experiments, we investigated the use of directed graphical models (DGMs), and found that when using non-lexical features, the DGMs outperform the hierarchical SVM classification method of Fernández et al. (2008b). The F1-score for the four DDA classes increased between .04 and .19 ($p < .005$), and for identifying decision discussion regions, by .05. This is encouraging because lexical features have disadvantages—for example they can be domain specific and greatly increase the feature space. In addition, modelling the dependencies between the DDA classes increased performance for *Agreement* utterances, and the DGMs were robust to ASR.

In the summarization experiments, we summarized decision discussions by extracting key words/phrases from their *Issue (I)* and *Resolution (R)* utterances. Each utterance's Word Confusion Network was parsed with an open-domain semantic parser, thus producing multiple candidate phrases, and then an SVM regression model selected one of these phrases to enter into the summary. The experiments here investigated the usefulness of a new SVM feature which measures the level of semantic similarity between candidate *I* parses and *R* utterances, and vice-versa. This feature was generated with a semantic-similarity metric which uses WordNet as a knowledge source. It was found to improve performance with both manual transcripts and ASR, and for *R* summarization, the improvements were highly significant ($p < .005$).

In future work, we plan to integrate lexical features into our DGMs by using a switching Dynamic Bayesian Network similar to that reported in (Ji and Bilmes, 2005). We also plan to extend the decision discussion annotation scheme so that we can try to automatically extract supporting arguments for decisions.

Acknowledgements This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-07-D-0185/0004, and by the Department of the Navy Office of Naval Research (ONR) under Grants No. N00014-05-1-0187 and N00014-09-1-0106. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or ONR.

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*.

John Dowding, Jean Mark Gawron, Doug Appelt, John Bear, Lynn Cherny, Robert Moore, and Douglas Moran. 1993. GEMINI: a natural language system for spoken-language understanding. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Raquel Fernández, Matthew Frampton, John Dowding, Anish Adukuzhiyil, Patrick Ehlen, and Stanley Peters. 2008a. Identifying relevant phrases to summarize decisions in spoken meetings. In *Proceedings of Interspeech*.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008b. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

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 *Proceedings of ICASSP*.

Pey-Yun Hsueh and Johanna Moore. 2007. 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*.

Gang Ji and Jeff Bilmes. 2005. Dialog act tagging using graphical models. In *Proceedings of ICASSP*.

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.

Claudia Leacock and Martin Chodorow, 1998. *WordNet: An Electronic Lexical Database*, chapter Combining local context and WordNet similarity for word sense identification. University of Chicago Press.

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, the 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, Netherlands.

Kevin Murphy. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California Berkeley.

Ted Pedersen. 2002. Semantic similarity package. http:/www.d.umn.edu/ tpederse/similarity.

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.

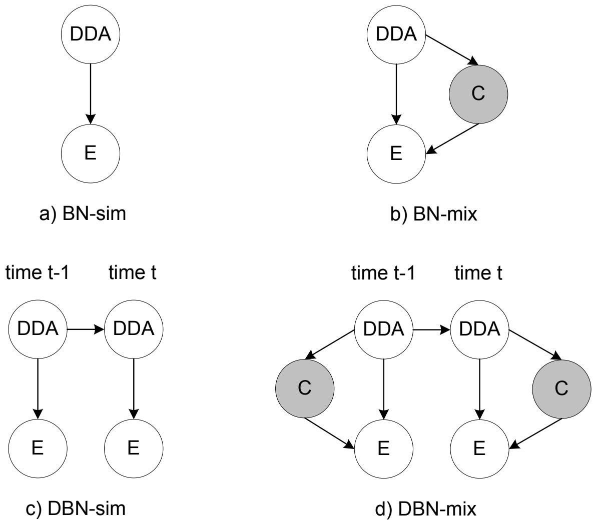
Andreas Stolcke, Xavier Anguera, Kofi Boakye, Özgür Çetin, Adam Janin, Matthew Magimai-Doss, Chuck Wooters, and Jing Zheng. 2008. The ICSI-SRI spring 2007 meeting and lecture recognition system. In *Proceedings of CLEAR 2007 and RT2007*.

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.

Steve Whittaker, Rachel Laban, and Simon Tucker. 2006. Analysing meeting records: An ethnographic study and technological implications. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 101–113. Springer.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Appendix



a) BN-sim

b) BN-mix

c) DBN-sim

d) DBN-mix

Figure 1: Simple DGMs for individual decision detection. During training, the shaded nodes are hidden, and the clear nodes are observable.

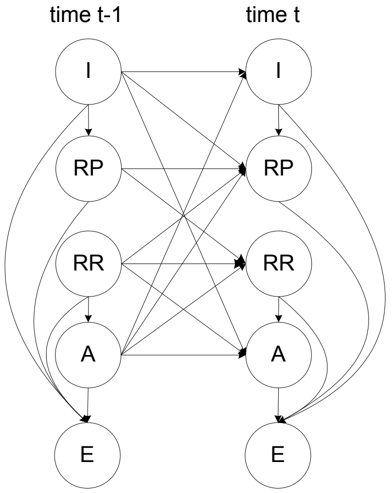


Figure 2: A DGM that takes the dependencies between decisions into account.