Detecting Arguing and Sentiment in Meetings

Swapna Somasundaran
Dept. of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
swapna@cs.pitt.edu

Josef Ruppenhofer
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
josefr@cs.pitt.edu

Janyce Wiebe
Dept. of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
wiebe@cs.pitt.edu

Abstract

This paper analyzes opinion categories like Sentiment and Arguing in meetings. We first annotate the categories manually. We then develop genre-specific lexicons using interesting function word combinations for detecting the opinions. We analyze relations between dialog structure information and opinion expression in context of multi-party discourse. Finally we show that classifiers using lexical and discourse knowledge have significant improvement over baseline.

1 Introduction

In this work, we bring together two areas of research which have seen great interest in recent times. Multi-party meetings have been analyzed with regard to dialog acts, hotspots, argumentation and decision points. Similarly, there is increasing activity in the automatic extraction of opinions, emotions, and sentiments in text (*subjectivity*) to provide tools and support for various NLP applications.

We believe that opinion information can enhance an interactive agent's ability to moderate a meeting; enable a summarizer to specifically report those opinions that influenced the decisions; and enhance the capabilities of Question Answering (QA) systems. As an example, consider a meeting from the AMI corpus (Carletta et al., 2005) where the participants have to design a new TV remote control. The following opinions are expressed regarding the TV remote:

U1. "It *[the remote]* is not as fast as a usual remote control"

U2. "That *[remote feature]* will be harder to learn"

U3. "We'll definitely won't go with that one *[speech recognition]*"

U4. "We can skip speech recognition directly, because it's not reachable for twenty five Euros".

Somebody who missed the meeting and had to find out details about the decisions made, may want to ask questions like:

Q1. "Why was the remote not rated highly?"

Q2. "Who argued against the speech recognition?"

Q3. "What were the points of persuasion against the speech recognition feature?"

Question *Q1* is answered by Utterances *U1* and *U2*, which express sentiments toward the remote. *Q2* is best answered by retrieving the names of all participants who had utterances similar to *U3* and *U4*. Similarly, *U4*, where the speaker is arguing for skipping the speech recognition would be a relevant answer for *Q3*. In order to be able to answer such questions, we explore two particular sub-types of subjectivity: Sentiment and Arguing. In the example utterances above, *U1* and *U2* express Sentiments, while *U3* and *U4* show speakers Arguing for their views. These subjectivity subtypes have proven useful for Question Answering on online multi-party debates (Somasundaran et al., 2007).

There has been a fair amount of work on the Sentiment category. By contrast, little work has been done on the Arguing category. We first define and annotate these opinion types in AMI meetings. We then perform inter-annotator agreement studies to verify if the two categories can be reliably detected.

We develop an Arguing lexicon as a new knowl-

edge source for automatically recognizing the Arguing category. We use previously developed lexicons for Sentiment detection (Wilson et al., 2005; Stone et al., 1966) to evaluate their portability to multi-party meetings. Previous efforts in recognizing opinions (or subjectivity) in monologic texts have focussed on knowledge from lexico-syntactic sources. While these have proven useful, we believe that in the conversational genre, reliably recognizing opinion expressions in utterances is a complex discourse task. Thus, we explore the novel use of dialog features for opinion recognition in combination with a lexicon. We find that this combination of knowledge sources shows promising results.

The rest of the paper is organized as follows: We introduce the data in Section 2 and our opinion definitions in Section 3. Then in Section 4 we present our annotation categories. In section 5 we explain the knowledge sources used for classification and present our experimental results in Section 6. Related work is discussed in Section 7 and finally we conclude in Section 8.

2 Data

For this work, we annotated 7 scenario-based team meetings from the AMI corpus resulting in a corpus of 4302 segments (6504 sentences) for our supervised learning experiments. In these meetings, four participants collaborate to design a new TV remote control in a series of four meetings, which represent different project phases, namely project kick-off, functional design, conceptual design, and detailed design.

In order to make the best use of the annotators' time in this work, we decided not to annotate the kick-off meetings as we believe them to be less rich in our opinion categories.

Each meeting in the AMI corpus comes with rich transcription and is annotated with dialog acts, argumentation, topics, etc. The corpus provides segment (turn) information for each speaker. Based on the rich transcriptions, we split the segments further into sentences. Sentence level classification tasks have a finer granularity and are of interest for applications like QA. On the other hand, in the absence of sentence boundary information, real time ASR systems work at the segment level. As there is interest at both levels of granularity, we present results at both the segment and sentence levels in this paper.

Some of the AMI annotations that are of interest in this work are Dialog Acts and their Adjacency Pairs. The AMI meeting is annotated with 15 Dialog Act (DA) categories: *Backchannel, Stall, Fragment, Inform, Elicit-Inform, Suggest, Offer, Elicit-Offer-Or-Suggestion, Assess, Elicit-Assessment, Comment-About-Understanding, Elicit-Comment-Understanding, Be-Positive, Be-Negative, Other*. Two DAs may be linked via an Adjacency Pair (AP) relation. One of the DAs is the source and the other is the target in the AP. There are 5 AP types, namely: *Support/Positive Assessment, Objection/Negative Assessment, Uncertain, Partial agreement/support, Elaboration*.

3 Opinion Definition

Our two opinion types are adapted from the work on attitude categories in monologic texts by Wilson et al (2005). They are defined as follows:

Sentiment: Sentiments include emotions, evaluations, judgments, feelings and stances. For example in the sentence "This idea is **good**", "good" expresses the sentiment.

Arguing: Arguing includes arguing for something, arguing that something is true, or should be done. Arguing brings out the participant's strong conviction and/or his attempt to convince others.

In multi-party discourse, speakers argue for something in a variety of ways. As arguing opinions are less well studied, we will examine some examples. Consider the following utterances, where the lexical anchors that indicate Arguing are shown in bold.

A1. "**I think** this idea will work"

A2. "This is the lightest remote **in the world**"

A3. "We **ought** to get this button"

A4. "**Clearly**, we cannot afford to use speech recognition"

A5. "It would be nice **if we could** have the curved shape"

A6. "I brought this up **because** this will affect the cost"

A7. "**We want** a fancy look and feel"

In *A1*, the speaker argues by explicitly stating his conviction. In *A2*, the speaker simply asserts his argument, while in *A3* the speaker argues for getting the button by framing it as a necessity. In *A4*, the speaker states his proposition categorically to argue

for it. Interestingly, in face to face conversations, participants also use persuasive constructs, justification or communal desire to argue for something as in *A5*, *A6* and *A7* respectively.

In examples *A1* to *A7* above, there are overt lexical anchors that indicate an arguing intent in the speaker's utterance. However, context, in addition to lexical clues is needed to infer that arguing is taking place. As part of a casual conversation, the utterance "I think John was at home" would not be Arguing, despite the presence of "I think". However, in a debate about John's whereabouts at the time of a murder, the sentence could function as Arguing. Here the context and the knowledge that there is a disparity between the speakers helps us infer that the sentence is intended to argue. Finally, sometimes arguing is done even in the absence of any overt lexical anchor. Consider:

A8. "The speech recognition is nice. Yes, speech recognition. It falls within our price range too"

In *A8*, we do not find any explicit markers. However, the speaker attempts to win approval for the speech recognition by his affirmation and his positive evaluations (sentiment) of the speech recognition and its price. These various elements together build up the argument.

4 Annotation Categories

Our annotation categories are Sentiment and Arguing. We discuss the varied ways of arguing in our annotation guide to help the annotators. As explained in Example *A8* of Section 3 sometimes Arguing is done without overt lexical anchors, which makes such cases difficult to annotate reliably. We assign these cases to a special category called Utterance Arguing.

We adapt the basic annotation frame for our opinion type from (Wilson and Wiebe, 2005). The relevant components of the frame are:

- **Text span:** The span of text that captures the opinion type. In the case of Utterance Arguing, this text span may cover the whole utterance.
- **Inferred: (true/ false)** This feature indicates that the annotator used inference for this annotation. For example, "very dark" is labeled as Sentiment in the sentence "This (TV) remote is **very dark**". This annotation is based on the knowledge that participants consider a dark color undesirable for the remote.
- **Annotator Confidence: (certain/ uncertain)** The annotators set this feature to uncertain when they are unsure of the annotation.

	Sentiment	Arguing	UtteranceArguing
segments	0.826	0.716	0.372
sentences	0.789	0.677	0.326
	Ignoring Annotator-uncertain cases		
segments	0.838	0.716	0.382
sentences	0.805	0.677	0.332
	Ignoring Annotator-uncertain and Inferred cases		
segments	0.85	0.716	0.382
sentences	0.814	0.677	0.332

Table 1: Kappa values for Inter-annotator agreement

4.1 Inter-annotator Agreement

Two annotators (two of the authors) underwent 3 rounds of training. Then we calculated inter-annotator agreement using Cohen's kappa over a previously unseen meeting (607 segments, 1002 sentences). Although the annotators tag expressions, agreement is calculated over the segment or the sentence. For this purpose, we assign a segment (or sentence) the labels of all the expressions annotated within it.

Table 1 shows the results of the agreement study. Our inter-annotator kappa values are in the Substantial Agreement Range according to Landis and Koch (1977) for Sentiment and Arguing, and in the Fair Agreement Range for Utterance Arguing. For Sentiment, when we exclude the labels from those instances that were tagged as inferential or uncertain, the agreement numbers go up to 0.85 for the segment and 0.814 for the sentence level respectively.

Compared to Sentiment, Arguing has lower kappa values at at 0.716 at the segment and 0.677 at the sentence level. We do not see any changes in the values when uncertain cases are removed. In this meeting the segments or sentence unit typically contain multiple expressions tagged for Arguing. Thus if an arguing label marked as uncertain was excluded from a given unit, but the unit had another label marked as certain elsewhere, then that unit overall still got an arguing label which counted toward the kappa calculation.

As expected, the Utterance Arguing category proved to be difficult. This is because it requires the annotators to infer whether the speaker is arguing when the utterance does not have any definite markers.

5 Knowledge Used in Classification

In this section, we discuss the development of our lexicon and the rationale for using dialog structures as knowledge sources for our automatic classifiers.

Much work in sentiment and subjectivity detection in monologic texts has focussed on lexical and syntactic features. In order to capture the lexical information we use lexicons. In the context of multiparty meetings, we hypothesize that the discourse flow and participant interaction act as useful indicators of opinion expression. We use Dialog Acts (DA) and Adjacency Pair (AP) features to capture the flow of discourse. We also believe that the lexical and discourse knowledge are complementary, and we build a system using all the features to test this hypothesis.

5.1 Sentiment Lexicon

We availed ourselves of previous work on Sentiment lexicon development, namely the General Inquirer (GI) (Stone et al., 1966), and Wilson et al's (2005) Subjectivity Clue list. The former provides a list of positive and negative words, while the latter contains a list of word and expressions that are strong/weak indicators of subjectivity, valence shifters, or intensifiers. In all, this gives us 6 lexicon categories to which a sentiment word may belong: GI Positive, GI Negative, Strong Subjective Clue, Weak Subjective Clue, Intensifier, and Valence Shifter.

5.2 Arguing Lexicon

We assembled an Arguing lexicon for meetings as follows. We inspected one AMI meeting (not used for training or testing) for words, phrases or word patterns that are indicative of Arguing. Then we explored the ICSI Meeting Recorder Dialog Act (MRDA) corpus (75 meetings, 72 hours) for similar expressions in order to develop more general patterns and increase the coverage of the lexicon. This was done in two steps. In the first, all instances of certain Dialog Act types *(dispreferred answer, negative answer, command, defending/explanation, suggestion)* were extracted and frequent n-grams $(1 \leq n \leq 4)$ identified. In the second phase, we manually inspected, for the highest ranking n-grams, a sample of 10-15 actual instances in the ICSI corpus and retained those n-grams that seemed promising. Finally, we looked over three ICSI transcripts in full to assess the coverage of the annotation concepts to be applied to the AMI data. This process produced a lexicon of 226 entries, sorted into 18 categories such as necessity, conditional, emphasis, generalization, contrast, causation, etc. to account for the various ways in which speakers argue.

Type	Example
emphasis	that's why the thing is
necessity	ought to had better
inconsistency	except that it's just that

Table 2: Examples from Arguing lexicon

As the entries given in Table 2 suggest, closed-class items such as modal verbs, adverbs, or conjunctions play a more important role in identifying instances of our Arguing class than open-class items. For instance, words like "oppose", "support", and "conclude" which directly denote aspects of arguing and reasoning are rare, whereas causal connectives such as "so", "because", and "if" are frequent.

We can understand the importance of closed-class items in terms of the distinction that Wiebe (2002) makes between direct subjective elements and expressive subjective elements. Direct subjective elements are exemplified, in the sentiment domain, by words like "love" or "criticize" which directly denote a particular kind of private state of a source, possibly in relation to a target, and which can realize their source and, if present, their target as a syntactic dependent. Expressive subjective elements, exemplified by words like "jerk" and "annoyingly", presuppose but do not denote a private state and cannot occur in syntactic construction with the source of the private state. Instead, the source is to be identified by the hearer from the candidate set made up by the interlocutors and the human referents in the discourse.

Applying this distinction to the Arguing category, we find that in the spoken conversation of meetings, where arguments are constructed in real-time, expressive subjective elements are prominent, with the

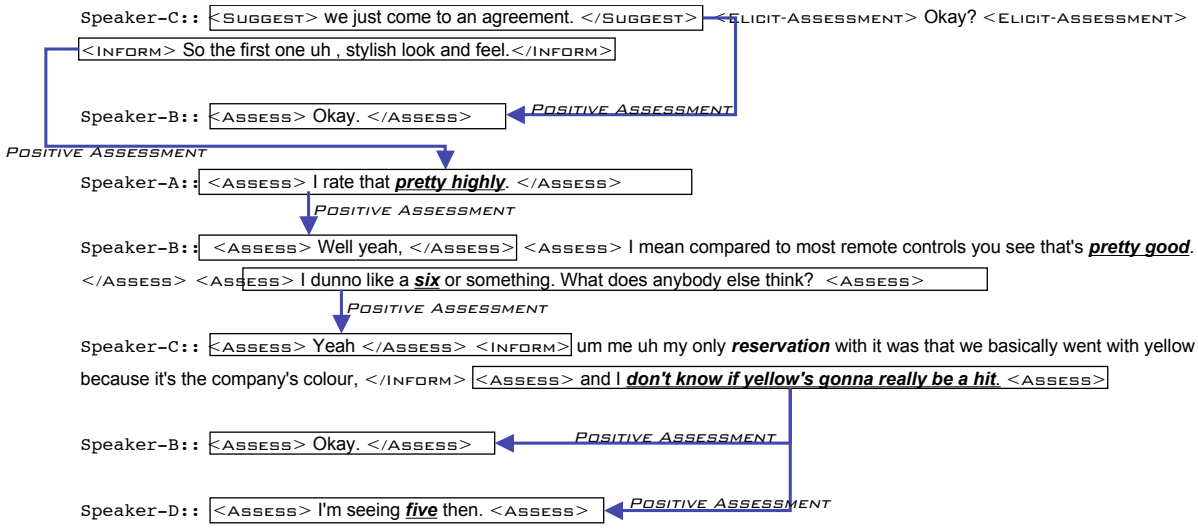

Figure 1: Sentiment expression and discourse flow.

sources typically being the speakers. This makes sense in particular for modal verbs such as "must", "need", etc. as arguing directly concerns modality: speakers discuss what is, what could be, what should be. By contrast, we find fewer direct subjective elements such as "require" or "argue". These elements, however, seem very suitable for reporting on arguments.

5.3 Dialog acts and Adjacency pairs

We observe that there is an interplay between our opinion categories and the dialog level annotations in the AMI corpus. Consider the following AMI meeting snippet where the participants rate their TV remote control design on a number of metrics such as learnability, look and feel, etc. using a scale from one (worst) to seven (best).

> *Speaker-C::* we just come to an agreement. Okay? So the first one uh , stylish look and feel .
> *Speaker-B::* Okay.
> *Speaker-A::* I rate that pretty highly.
> *Speaker-B::* Well yeah, I mean compared to most remote controls you see that's pretty good. I dunno like a six or something. What does anybody else think?
> *Speaker-C::* Yeah um me uh my only reservation with it was that we basically went with yellow because it's the company's colour , and I don't know if yellow's gonna really be a hit.
> *Speaker-B::* Okay.
> *Speaker-D::* I'm seeing five then.

Figure 1 illustrates the opinion annotations (in bold underlined text spans), DA annotations (as enclosing XML tags) and AP annotations (as directed links between segments) of the above meeting snippet. *C* introduces the first metric for evaluation, the stylish look and feel. *A* has a positive *Sentiment* about the remote in this regard and hence says he rates it "pretty highly". *B* shares *A's* positive *Sentiment*. He too evaluates the remote favorably and judges it as deserving a rating of six. Note that here "six" is considered an inferred sentiment, as it reflects the participant's evaluation of the remote. *C*, however, shows his negative *Sentiment* towards the remote by pointing out his reservation about the choice of the color yellow. *C's Sentiments* convince *D*, who then evaluates the look and feel at the lower grade of 5.

The Dialog Acts and Adjacency Pairs that capture the exchanges between the participants are indicative of the Sentiments expressed. For example, it is likely that a participant who has a positive evaluation of an object might positively assess his preceding speaker's positive assessment of the same object. We see this in Figure 1 when *A* and *B* both show positive *Sentiment* towards the remote's look and feel. *B* shows a *Positive Assessment* of *A's Assessment*. *D*, who evaluates the look and feel of the remote at a lower grade (negative *Sentiment*) has a *Positive Assessment* toward *C's Assessment* (negative *Sentiment*) of the remote. Thus the participants' sentiments towards objects are also reflected in their interpersonal dialog acts and vice versa. We also

found interesting relations bettween arguing and dialog structure. Due to space considerations, this is discussed in Appendix A

We believe Dialog structure (DA and AP) and our opinion categories are complementary rather than interchangeable. Dialog acts are focused on interpersonal exchanges and discourse functions, while opinion categories are focused on participants' private states usually towards objects (which may be other participants). In our corpus we found that it is not always necessary for a Sentiment instance to be associated with an Assess Dialog Act. Consider the utterance: "Okay, so when you have a lot of room inside. So you can make it **very easy to use**. **'Cause** you can write a lot of comments besides it." This sentence was labeled as an *Inform* DA as it functions to inform the participants of the roomy interior of the remote control. Orthogonally, it was tagged as a positive *Sentiment* ("very easy to use") and positive *Arguing* ("'Cause").

6 Experiments and Results

In this section, we perform machine learning experiments to test our hypothesis that our knowledge sources from Section 5 are useful. We perform supervised machine learning on our annotated corpus of 4302 segments (6504 sentences) using a standard SVM package (Joachims, 1999). The recognition of each opinion category is formulated as a binary classification problem. We do not attempt automatic classification for Utterance Arguing as we consider our inter-annotator agreement for this category to be too low to form a reliable gold standard.

We use two baselines: a majority-class dumb baseline that guesses false every time, and a smart SVM classifier trained on a bag of words (BOW). Then we add our opinion features individually or in combination to the baseline classifier. The lexicon features for the BOW+lex classifier are counts of words from each lexicon type in the given segment or sentence.

The AMI DA types introduced in Section 2 form the additional features for the BOW+DA classifier. The AP links described in Section 2 along with their source DA and target DA form a DA-AP-DA chain. These DA-AP-DA chains form the features for the BOW+AP classifier. Since we do not make a polarity distinction, we conflate *Positive* and *Negative Assessment* into a single category *Assessment*. As there are 15 DAs and 4 APs (after conflation) there are $15 * 4 * 15 = 900$ possible combinations; however of these, only 99 types actually occur in our annotated corpus. The BOW+DA+AP classifier has all the DA features and the AP features; the BOW+All classifier uses DA, AP and lex features.

	Acc	Prec	Rec	F-measure
	Segment Level classification			
BOW	88.42	69.52	51.95	57.99
BOW+lex	88.84	70.1	**53.07**	**59.16**
BOW+DA	89.28	**73.81**	**54.62**	**61.26**
BOW+AP	88.73	70.1	53.07	59.16
BOW+DA+AP	89.24	**73.14**	54.38	**60.9**
BOW+All	89.28	**73.17**	**54.98**	**61.37**
	Sentence Level classification			
BOW	89.43	69.22	46.69	54.62
BOW+lex	89.51	69.12	48.04	55.53
BOW+DA	89.80	71.11	49.07	**56.7**
BOW+AP	89.40	69.42	46.21	54.11
BOW+DA+AP	89.79	71.29	48.87	56.54
BOW+All	90.3	**73.22**	**51.32**	**59.20**

Table 3: Arguing Classification Results.

The accuracy of the majority-class Arguing classifier is 82.84% at the segment and 85.5% at the sentence level. All the classifiers, including the smart baseline (BOW), improve over this by about 7 percentage points at the segment and by about 4 percentage points at the sentence level. Table 3 shows the performance of our Arguing classifiers. All results are reported over 20-fold cross-validation. The results that are significantly better ($p < 0.05$) than the smart BOW baseline are shown in bold. The results in Table 3 indicate that the DA features are useful for detection of Arguing. The only classifier that performs significantly better than the smart baseline at both the segment and sentence level is the one that uses all the features (BOW+all). This corroborates our hypothesis that lexical and discourse information are complementary. The Arguing lexicon significantly improves recall and f-measure for segments, but the results are not significant at the sentence level. We think this is because our preliminary lexicon with its lesser coverage can still succeed in finding matches in the larger segmental units, but fails in the smaller sentential units. We believe increasing the breadth of coverage will remedy this. Table 4 shows the performance of the Sentiment

	Acc	Prec	Rec	F-measure
	Segment Level classification			
BOW	86.87	80.84	48.53	58.77
BOW+lex	88.29	81.43	**56.14**	**65.18**
BOW+DA	87.45	81.93	**51.48**	**62.0**
BOW+AP	87.27	81.02	**50.69**	**61.11**
BOW+DA+AP	87.36	82.73	49.55	**60.93**
BOW+All	88.66	82.01	**57.89**	**66.88**
	Sentence Level classification			
BOW	88.23	82.41	44.08	56.61
BOW+lex	89.77	81.99	**54.70**	**64.89**
BOW+DA	88.59	82.11	**47.08**	**59.14**
BOW+AP	88.67	82.68	**46.73**	**58.97**
BOW+DA+AP	88.64	82.47	**47.1**	**59.22**
BOW+All	89.95	82.49	**55.42**	**65.62**

Table 4: Sentiment Classification Results

classifiers. Here too, using all features gives the best performance at both the segment and sentence level. Additionally, we also see that each of our features, lexical or dialog-based, individually improve the recall and f-measure. The accuracy of the majority-class Sentiment classifier is 79.12% at the segment and 82.16% at the sentence level. The best classifier (BOW+All) improves over this by about 9 percentage points at the segment and 8 percentage points at the sentence level. We also see that the lexicons from the monologue text genres help in improving the recall significantly. It is encouraging that resources developed for extracting sentiments from monologic texts will be useful for processing conversational data as well.

7 Related Work

Sentiment detection is being carried out across a variety of genres and at various levels (e.g. document level by Thomas et al. (2006), phrase level by Wilson et al. (2005)).

Like much other work on subjectivity (e.g. Nasukawa and Yi (2003)), we use lexicons as knowledge sources in classification. Somasundaran et al. (2007) use a lexicon for detecting Arguing in text. In contrast, our work is on multi-speaker conversations. Biber (1988) in work on textual variation identifies a dimension of "Overt persuasion" whose categories (e.g. modal verbs and conditionals) are similar to the expressions we gathered in our lexicon. Ducrot (1973) studies arguing related items, but his work is on French and is not corpus-based. A vast body of work exists within linguistics, rhetoric and philosophy that is relevant to arguing (e.g.(Dancygier, 2006; van Eemeren and Grootendorst, 2004)).

With regard to meetings, the most closely related work includes the dialog-related annotation schemes for various available corpora of conversation (Dhillon et al. (2003) for ICSI MRDA; Carletta et al. (2005) for AMI; Burger et al. (2002) for ISL). We think our annotation scheme complements the annotations provided in these corpora in that it adds finer granularity for statement-speech acts by distinguishing expressions of sentiment and arguing from objective statements.

Our work also connects to research on hot spots (Wrede and Shriberg, 2003), and efforts to annotate the mental states of participants in meetings or interviews on the basis of multi-modal data (Devillers et al., 2005; Reidsma et al., 2006). The focus of these kinds of research is different from ours in that they target the actual mental states of the speakers in the unfolding situation, while we focus on subjective states communicated through language. While often the same, they are not necessarily identical as language allows for displacement: participants may calmly report about other people's anger, report their past or expected future mental states, etc. Our approach is similar to the one used by Galley et al. (2004) where adjacency pair information is used to detect agreement/disagreement amongst participants. Similarly, in the prediction of congressional vote, Tomas et al. (2006) use adjacency pair information to detect agreement amongst speakers. Another closely related area is argument diagramming of meetings (Rienks et al., 2005), where lines of deliberation are analyzed without making a subjective/objective distinction. Our work can also be combined with ongoing work on decision detection (Hsueh and Moore, 2007; Purver et al., 2006). While our annotations track opinions in the decision making process, the decision detection research is mostly concerned with its outcome.

8 Conclusion and Future Work

We presented the annotation of the Opinion types Sentiment and Arguing on meetings. We developed a new lexical resource for the Arguing category. We showed that previously developed Sentiment lexi-

cons have good coverage in the new genre. We hypothesized that dialog structure interacts with the expression of opinions and confirmed this through machine learning experiments. Finally, using all the features gave the best performance, confirming our hypothesis that both lexical and discourse information is needed to detect opinions in multi-party conversations.

Our future work will involve increasing the breadth and reliability of our arguing lexicon both manually and via automatic means. We also plan to use richer discourse and meeting level information as well as study interactions between opinion types.

References

D. Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press.

S. Burger, V. MacLaren, and H. Yu. 2002. The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. In *ICSLP 2002*.

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 Meetings Corpus. In *Proceedings of the Measuring Behavior Symposium on "Annotating and measuring Meeting Behavior"*.

B. Dancygier. 2006. *Conditionals and Prediction: Time, Knowledge and Causation in Conditional Constructions*. Cambridge University Press.

L. Devillers, S. Abrilian, and J.-C. Martin. 2005. Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. In *Proc. of ACII*.

R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. 2003. Meeting recorder project: Dialog act labeling guide. Technical report, ICSI Tech Report TR-04-002.

O. Ducrot. 1973. *Le preuve et le dire*. Mame.

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *ACL*.

P.-Y. Hsueh and J. Moore. 2007. What decisions have you made: Automatic decision detection in conversational speech. In *NAACL/HLT 2007*.

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

R. Landis and G. Koch. 1977. The measurement of observer agreement for categorical data. In *Biometrics, Vol. 33, No. 1*.

T. Nasukawa and J. Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *K-CAP 2003*.

M. Purver, P. Ehlen, and J. Niekrasz. 2006. Detecting action items in multi-party meetings: Annotation and initial experiments. In *Machine Learning for Multimodal Interaction*. Springer-Verlag.

D. Reidsma, D. Heylen, and R. Ordelman. 2006. Annotating emotions in meetings. In *LREC 2006*.

R. Rienks, D. Heylen, and E. van der Weijden. 2005. Argument diagramming of meeting conversations. In Vinciarelli A. and Odobez J., editors, *Multimodal Multiparty Meeting Processing, Workshop at the 7th Intl. Conference on Multimodal Interfaces*.

S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov. 2007. QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *Intl. Conference on Weblogs and Social Media*.

P. J. Stone, D. Dunphy, M. S. Smith, and D.M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

M. Thomas, B. Pang, and L. Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.

F. H. van Eemeren and R. Grootendorst. 2004. *A systematic theory of argumentation*. Cambridge University Press.

J. Wiebe. 2002. Instructions for annotating opinions in newspaper articles. Dept. of Computer Science Technical Report TR-02-101, University of Pittsburgh.

T. Wilson and J. Wiebe. 2005. Annotating attributions and private states. In *Proc. of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

T. Wilson, J. Wiebe, and P Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005*.

B. Wrede and E. Shriberg. 2003. Spotting hotspots in meetings: Human judgments and prosodic cues. In *Eurospeech*.

A Appendix A. Arguing Opinions and Discourse Flow

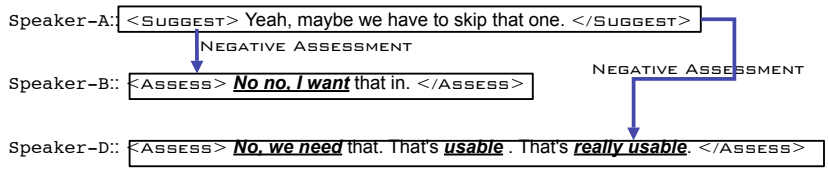

Figure 2: Arguing expression and discourse flow

As with the Sentiment opinions, for the Arguing category, too, we found an interrelation with Dialog Act exchanges. Consider the AMI meeting snippet below where the participants are discussing a beeping functionality. Speaker A has just suggested skipping it.

> *Speaker-A::* Yeah, maybe we have to skip that one.
> *Speaker-B::* No no, I want that in.
> *Speaker-D::* No, we need that. That's usable . That's really usable.

Figure 2 illustrates the annotations on this snippet. *A Suggests* that they might skip the beeping functionality. *B Argues* against this *Suggestion* with a vehement "No no". The "I want" in *B's* utterance acts as both *Sentiment* (positive towards the thing wanted) as well as *Arguing*. Thus, there is a *Negative Assessment* link between the two. *D*, too, *Argues* against *A's Suggestion* by stating that the beeping functionality is a necessity. He justifies this stance by evaluating the remote as usable and then reinforces his argument though repetition and intensification.