Towards a Representation for Understanding the Structure of Multiparty Conversations

Michael Ellsworth[†*]

TR-11-003

November 2011

Abstract

Dialog is a crucially important mode of communication. Understanding its structure is vital to a great number of technological innovations. Within the domain of meeting dialog applications alone, there is a need for navigating, summarizing, and extracting action items in recorded meetings and automated facilitation of meetings. These efforts, however, are hampered by a lack of agreement on how best to annotate meeting dialog to serve downstream applications.

In this technical report, I describe of some of the efforts at representation and annotation of dialog acts, surveying some of the vast differences between approaches. In addition, in order to more directly compare different representations of discourse structure, two small, pilot annotations were carried out on a subset of the ICSI Meeting Corpus using different annotation schemes. The results support the idea that skew between different annotation systems renders them to some degree incompatible.

† ICSI, 1947 Center St., Ste. 600, Berkeley, CA 94704
* UC Berkeley, Berkeley, CA 94720

This work was partially supported by funding provided to ICSI through National Science Foundation grant CI-P: 0958561 (“Towards a Consensus Representation for Understanding Structure of Multiparty Conversations”). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of the National Science Foundation.

Meetings are an extremely important part of modern life, affecting everything from hiring and firing decisions to which patients are given surgery [4]. And yet there is no general, efficient mechanism for recording meetings so that we can browse them, track the status of discussion items later, or effectively and efficiently represent the structure of meetings for human browsing or automatic meeting processing. Most meetings have no recording except for manually written subjective notes about action items and decisions. While there are a number of studies about representing and detecting events in multiparty conversations, such as action items, agreements/disagreements, and decisions, the spoken language research community has not yet reached a consensus on how to represent meeting structure.

In this report, I summarize previous efforts and corpora, additionally describing two small pilot annotations of the ICSI meeting corpus, first with a tagset designed for the AMI corpus (TAS) and second with a simple partition into questions, statements, and other. The overall goals of the annotation are to help in identifying points of agreement and problem areas across differing annotation schemes.

1 Models of Dialog Acts

There are a large number of models that have been used to describe spoken language. Some of these have only been used for linguistic description, while others have been used for annotation. A smaller subset has been used to represent meetings. In this section I compare a cross-section of these theories.

1.1 LDM (Linguistic Discourse Model) [20][21]

A syntactically-informed theory of discourse based on the notion of grouping basic discourse units (BDUs) into discourse constituency units (DCUs) by using relations to bind new DCUs or BDUs to the right edge of the graph. This theory was elaborated to form the Unified Linguistic Discourse Model (U-LDM).

1.2 RST (Rhetorical Structure Theory) [18][19]

This theory assumes a small, but indefinite number of different relations, with no claim of cross-linguistic identity. Several different partitions of the relations are possible, depending on need of the analyst. Relations are generally asymmetrical, differentiating nodes as “nucleus” and “satellite”, but

some relations involve only multiple nucleii. In any case, the output of analysis is a tree with labeled relations between the branches.

The RST Corpus consists of 385 WSJ articles (176k text) available through the LDC.[8] Although the text-type is widely divergent from that of meetings, the annotation shows that RST is useful for annotating natural language, and there is no clear reason why it could not be used for annotating meetings as RST has also been used for diverse kinds of text generation.[22]

1.3 DRT (Discourse Representation Theory) [16] and SDRT (Segmented Discourse Representation Theory) [2] [17]

In this theory, a discourse representation (DRS) is an ordered pair mapping from sets of entities to sets of intensional-semantics predications that govern those entities.

SDRT is based on DRT, differing primarily in including discourse elements including events and discourse chunks, with intensional-semantics predicates involving those entities. The predicates used to describe discourse relations do not form a delimited set in this theory, but they certainly include familiar relations like Elaboration.

Advocates of SDRT over RST cite its richer graph structure as a significant modeling advancement over the strict trees of RST.[10]

1.4 DAMSL (Dialog Act Markup in Several Layers) and SWBD-DAMSL

The DAMSL annotation paradigm was designed with practical annotation efforts in mind. It went through several iterations of development, ending up with a very complex system that separated out 15 different independent dimensions that describe dialog acts in different ways, e.g. attempt to influence addressee vs. signaling agreement vs. relation to the topic under discussion. While separating out these dimensions is quite attractive in terms of being able to use the information to reason about the dialog act in downstream tasks, "it may be difficult to train annotators to use a 15 dimension framework and data analysis becomes more difficult. You should keep this in mind if you chose to use the dat tool." [1]

SWBD-DAMSL is a scheme that was used to annotate the Switchboard Corpus. Although it might be considered, in some ways, a derivative of DAMSL, and there is a published mapping [15], the tagset is considerably different, including elaboration of question and answer types, social func-

tions, and statement presentation types like hedging. The tagset is not separated into the dimensions described for DAMSL, although multiple simultaneous tags were allowed and the mapping onto DAMSL tags allows most of the tags to be classified under the DAMSL dimensions.

1.5 MRDA (Meeting Recorder Dialog Act) [14][11]

The MRDA annotation scheme was developed from Switchboard DAMSL (see mapping in MRDA-manual, 1.2), but some functions described for S-DAMSL are not annotated (including categorizing dialog acts as openings, closings, or discussions of communication), and some that were unannotated in S-DAMSL are annotated in MRDA (including categorization as floor-holders and floor-grabbers).

1.6 MALTUS (Multidimensional Abstract Layered Tagset for UtteranceS) [9]

This effort is a more constrained derivative of MRDA and Switchboard tagsets designed in the course of the IM(2) project to be easier for automatic annotation.

1.7 MRMA (Meeting Recorder Meeting Act) [3]

This project focused more specifically on labeling how meeting participant interactions furthered meeting discussions and decision making. The annotation consists largely of detailed descriptions of what subprocess of a meeting an utterance belongs to (administrative overhead, decision making, discussion, interruptions, humor, and their subtypes) and indications of agreement, number of participants, time-scale of plans, and other information related to what occurs in meetings.

This annotation framework is quite divergent from the other efforts described here in that it focuses on how participants' interaction fits into a specific schema of how meetings work, and is thus both richer in the kind of information one might want to extract about a meeting and, necessarily, not directly relevant for annotation of other genres.

1.8 DIT++ (Dynamic Interpretation Theory) [6][5]

DIT++ is a proposed universal theory of discourse acts, loosely similar in its multidimensionality to DAMSL. In addition to the specific dimensions, however, it has a defined set of general-purpose communicative functions

(e.g. statement vs. question) that can apply at any of the levels. The system has been applied to a subset of the AMI, OVIS, and DIAMOND corpora [13] and a set of Dutch-language corpora. The most recent release has developed in conjunction with a proposed ISO standard for dialog act annotation.[7]

1.9 TAS (Twente Argumentation Schema)[23]

The AMI and AMIDA projects introduced a new annotation methodology focused on capturing the structure of argumentation in meetings. It was based conceptually on the long tradition of Wigmore's charting method and the Toulmin Model of Argumentation. Since this model is focused on argumentation, it does record most of the dimensions of dialog acts considered in other models, such as floor management, interruption, self-correction, communication mode (e.g. humor) or performing socially required speech acts, like apologies, introductions, and leave-taking. Instead, it focuses on how statements involved in a decision-making task are related to each other.

2 Pilot TAS Annotation of the ICSI Meeting Data

In order to see how alternative annotation systems compare to each other, we decided to annotate a small portion of the ICSI Meeting Corpus, which was already annotated with ICSI-MRDA annotation, with TAS-style annotation. Since our interest was primarily on the *structure* of meetings, TAS's lack of multidimensionality made it ideal for our purposes.

2.1 (Minor) Changes in the TAS System

Because we were most interested in how meetings are structured, we did not focus on changing the dialog act tags or the relation labels, but rather focused on relaxing the constraints imposed in the original TAS annotation. In contrast to the TAS manual's directions to annotate single turns, annotators were allowed to annotate units larger or smaller than the pre-set turns. In addition, we removed the the manual's constraint to add dependent relations to dialog acts only in such a way that the relation lines do not cross.

2.2 Results and Questions from the Pilot

Two annotators were given a small set of meetings from the ICSI Meeting Corpus to annotate. Even using a very loose criterion (annotation overlap), inter-annotator agreement (table 1) on dialog acts is modest. This is not surprising given the small scale of the task and the concommitant short period of training that annotators received. When TAS tags are mapped to MRDA tags as annotated in the Meeting Recorder Project (MRP)[1], agreement of each annotator with the MRP is similar, and both are similar to inter-annotator agreement (see table 1). The slightly higher average agreement between annotators (75.2%) as compared with their agreement with MRP (74.6%) is likely due to a slight skew in the mapping of TAS tags to MRDA tags attributable to the difference in what the tagging systems are meant to represent. For example, statements like "We need to talk about anonymization" are classified as *Open Issue* in TAS, but in the MRDA tagset are labeled *Statement*, which naturally maps onto TAS *Statement*.[2] Implicit in the agreement achieved with the MRP annotation for dialog acts is the fact that MRDA annotation can be automatically transduced into TAS annotation at approximately the same accuracy (75%) as the (admittedly briefly trained) human annotators on this small project.

No attempt is made here to systematically compare MRP adjacency pair annotation with relation annotation, as, unlike TAS relations, MRP adjacency pairs are not directional and generally connect only isolated pairs of dialog acts rather than connecting all parts of a meeting into a single graph. Only 40% of MRP dialog acts have an associated adjacency pair annotation, whereas the TAS annotation specs demand that nearly 100% of acts have relation information (97% in the TAS annotation here). As a result, the data are essentially incomparable. In addition, since interannotator agreement is quite low (see last two rows of table 1), the data would not support any meaningful comparison in any case.

In sum, while the present pilot annotation supports the idea that dialog act annotation in different annotation schemes is comparable and, to some extent, interconvertable, the fundamentally different principles of relation geometry prevent comparability of MRDA adjacency pairs and TAS relations.

[1] See http://icsi.berkeley.edu/∼infinity/TAS-MRDAmapping.txt.

[2] Inter-annotator agreement is not strictly comparable to the MRDA project's reported Kappa statistics (.75 for a six-way distinction), but surely represents lower overall agreement.[12]

	1 vs. 2	1 vs. MRP	2 vs. MRP
All dialog acts	75.2%	74.5%	74.6%
Bed004	77.6%	78.3%	73.3%
Bed005	80.1%	73.8%	80.5%
Bns003	71.2%	73.2%	69.3%
Bro005	70.4%	73.8%	69.3%
Typed relations	45.7%	—	—
Untyped relation	65.7%	—	—

Table 1: Agreement for pilot TAS annotation of sections of the ICSI Meetings Corpus. Agreement numbers for two annotators compared to each other and to the Meeting Recorder Project's MRDA tags mapped to TAS dialog acts. Total TAS annotation: 4301 dialog acts and 4203 relation annotations over two annotators.

3 Second Annotation of ICSI Meeting Data

As a result of discussion with interested members of the community, it was decided that the project would take on another pilot annotation project with the goal of distinguishing between the effect of context and audio in providing cues for question/statement disambiguation. Two annotators were presented with pieces of the ICSI Meeting Corpus to annotate under three different conditions: A. transcript of isolated utterance only, B. transcript of utterance with two preceding and following utterances, and C. audio and transcript of utterance with audio and transcript context. To keep the task simple, the annotation scheme consisted of four labels: Statement, Question, Other, and Unclear. 'Statement' and 'Question' are defined (similarly to TAS) in terms of conversational intent rather than intonation or construction. 'Other' was used for cases like directives and conversation management. 'Unclear' was used for fragmentary utterances; when this tag was used by either annotator, the utterance was excluded from analysis. Because we were specifically interested in how annotators' knowledge was influenced by condition, we also asked annotators for confidence levels (1-10) for each annotation.[3]

Due to the briefness of the project, annotators were only able to annotate

[3]No statistically significant effects were found in the confidence data, so it is not discussed further here.

	Agreement (1 vs. 2)	vs. task C		vs. MRP	
		1	2	1	2
Task A	84.3% (1590)	79% (19)	78% (23)	75.0% (821)	80.6% (1759)
Task B	83.5% (3299)	77% (31)	71% (24)	74.3% (1729)	80.5% (1812)
Task C	79.0% (244)	—	—	65.6% (256)	69.4% (271)

Table 2: Percent agreement within and across tasks. Column 1: agreement between annotators within each task; columns 2a, 2b: each annotator's agreement with their own annotation of task C; columns 3a, 3b: each annotator's agreement with the Meeting Recorder Project annotation. (Total number of annotations compared in parenthesis).

	Statement	Question	Other
1 task A	82%	11%	7%
1 task B	77%	10%	13%
1 task C	70%	16%	14%
2 task A	90%	10%	0.3%
2 task B	86%	8.5%	5.4%
2 task C	76%	12%	12%
Mapped MRP	75%	6.9%	18%

Table 3: Percentage of dialog act types per annotator and task compared to the Meeting Recorder Project annotation of questions and statements.

a small subset of dialog acts in the Meeting Corpus. These were randomly chosen. In order to avoid annotators using the order of presentation in some way, the utterances were presented in different random orders in task A (utterance transcript only) and B (transcript with context), while in task C, utterances were presented in the order of the original transcript. In table 2 is shown the agreement numbers for each annotator against each other (column 1), versus their own annotation on task C (column 2), and versus the annotation of the MRP mapped onto these simple categories. Table 3 shows the frequency of each label. Given these frequencies, interannotator agreement is notably low.

Although the sample size of task C is far too small to draw any firm conclusions[4], there is definitely a trend in table 3 towards annotators over-

[4]This is due to the fact that in task C, the annotators used the NXT annotation tool,

identifying utterances as statements, which decreases with more information. Annotator 1's distribution of frequencies approximates that of the expert annotation of MRP with the contextual information of task B, while annotator 2's distribution is only similar to that of the MRP in task C, with audio and context. The changes in frequency skew along the tasks also explains, to some extent, the seemingly paradoxical fall in agreement numbers from tasks A to B to C (see table 2). Since the frequencies of the classes are more balanced in tasks B and C, agreement by chance is considerable rarer in tasks B and, especially, task C.

In sum, the propensity of annotators for choosing the most common class is greater in condition A than in B or C. This fact is consistent with annotators choosing the most common class when they have less information and less certainty (though as noted above, this did not show up in the confidence scores).

4 Discussion

The annotation tasks in this project were too small to prove any particular theses, but it is hoped that the results are suggestive of further research. Not surprisingly, there are multiple lines of evidence in the data above showing that comparability between systems (e.g. MRDA vs. TAS or the simple Statement/Question/Other annotation) is lower than interannotator agreement. This divergence is due in part to fundamental incompatibilities even in annotation schemes used on the same data and is not likely to improve when we compare more dialog tagging systems, as the models that have been used for annotation vary widely in their structure and focus and there are several more dialog-act and rhetorical-structure models that have not been tried out in significant annotation tasks. As can be seen from this diversity, and as was apparent in a workshop held in association with this work, it is not clear that there is any consensus on what features of dialog act systems are most useful, beyond the fact that most researchers believe that dialog annotation systems should be various to conform with the uses that the annotation is meant to be put to.

with full meeting audio and transcript, which was much slower than annotating a text file as in tasks A and B.

References

[1] James Allen and Mark Core. Draft of damsl: Dialog act markup in several layers: http://www.cs.rochester.edu/research/speech/damsl/revisedmanual/.

[2] Nicholas Asher. *Reference to Abstract Objects in Discourse*. Kluwer, 1993.

[3] Rebecca Bates, Elizabeth Willingham, Chad Kuyper, and Patrick Menning. Mapping Meetings Project: Group Interaction Labeling Guide (Draft) http://bates.cs.mnsu.edu/mrma.pdf.

[4] E. Boyd. Bureaucratic authority in the company of equals: The interactional management of medical peer review. *American Sociological Review*, 63(2):200–24, 1998.

[5] Harry Bunt. DIT++ Website: http://dit.uvt.nl/.

[6] Harry Bunt. Dimensions in Dialogue Act Annotation. In *Proceedings of LREC 5*, Genoa, 2006.

[7] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. Towards and ISO standard for dialogue act annotation. In *Proceedings of LREC 7*, Malta, 2010.

[8] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Denmark, 2001.

[9] Alexander Clark and Andrei Popescu-Belis. Multi-level Dialog Act Tags. In *Proceedings of the 5th SIGdial Workshop at NAACL*, 2004.

[10] Laurence Danlos, Bertrand Gaiffe, and Laurent Roussarie. Document sructuring a la sdrt. In *Proceedings of the European Workshop on Generation, ACL*, Toulouse, 2001.

[11] Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. Meeting Recorder Project: Dialog Act Labeling Guide. Technical report, ICSI, Berkeley, CA, 2004.

[12] Elizabeth Shriberd et al. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the HLT-NAACL SIGDIAL Workshop*, pages 97–100, Boston, 2004.

[13] Jeroen Geertzen, Volha Petukhova, and Harry Bunt. Evaluating Dialogue Act Tagging with Naive and Expert Annotators. In *Proceedings of LREC 6*, Marrakech, 2008.

[14] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The ICSI Meeting Corpus. In *Proceedings of ICASSP*, 2003.

[15] Dan Jurafsky, Liz Shriberg, and Debra Biasca. Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13: http://www.stanford.edu/ jurafsky/ws97/manual.august1.html, 1997.

[16] Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer, 1993.

[17] Alex Lascarides and Nicholas Asher. *Logics of Conversation*. Cambridge University Press, 2003.

[18] William Mann and Sandra Thompson. Rhetorical Structure Theory: description and construction of text structures. Technical Report ISI/RS-86-174, Information Sciences Institute, Nijmegen, The Netherlands, 1986.

[19] William Mann and Sandra Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[20] Livia Polanyi. The Linguistic Discourse Model: Towards a Formal Theory of Discourse Structure. Technical report, BBN Labs, Cambridge, MA, 1986.

[21] Livia Polanyi, Christopher Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. Sentential Structure and Discourse Parsing. In *Proceedings of the Discourse Annotation Workshop at ACL*, 2004.

[22] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.

[23] Rutger Rienks and Daan Verbree. TAS Annotation Manual: http://mmm.idiap.ch/private/ami/annotation/tas-annotation-manual.pdf.