

An Annotation Scheme for Agreement Analysis

Siew Leng Toh, Fan Yang, Peter A. Heeman

Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Science & Engineering University

siewlengtoh@yahoo.com fly@cslu.ogi.edu heeman@cslu.ogi.edu

Abstract

To accomplish a task that requires collaboration, people would first agree on a strategy and then together carry it out [1]. Our research interest lies in understanding how people explore different strategies and reach an agreement in conversation. We began by examining two-person dialogues in a very limited domain, in which we could just focus on the agreement process. In this paper, we describe an annotation scheme of coding the conversants' behaviors of exploring possible strategies, suggesting and accepting the optimal one, and then maintaining it. We report the inter-coder reliability of the annotation scheme on three expert annotators and two non-experts.

Index Terms: annotation scheme, agreement, inter-coder reliability, dialogue.

1. Introduction

There is a lot of interest in the automatic analysis of multi-party meetings, for summarizing what happened, determining what action items were assigned [2], and for building computer agents that can participate in them. A prerequisite for such work is to collect a corpus of these conversations, transcribe them, and then annotate them as to what is going on. Typically, a speech act scheme, such as DAMSL [10], might be used, which captures what the individual utterance is doing, such as asking a question, making a statement, or acknowledging what was said [3]. However, often times, we want to know what the utterance is about, what role it is playing in the dialogue. Such an annotation scheme might classify which utterances are summarizing a plan, suggesting a plan, or repairing it [4].

Multi-party conversations can be very difficult to analyze, especially when the topics under discussion are wide ranging, and vary from conversation to conversation (e.g. [5]). This makes it difficult to determine what each utterance is about. Furthermore, in order to capture the wide range of behaviors presented in the corpus, a complex annotation scheme is needed.

In this paper, we propose taking a more modest approach. As a first step, we examine two-person conversations on a very limited domain, that of playing a collaborative card game, in which the two players work together to assemble a poker hand [6]. These conversations incorporate information sharing, exploring different strategies, suggesting strategies, and maintaining the strategy. Furthermore, these conversations also require participants to repeatedly re-examine whether a strategy is still optimal, as new cards are dealt. Hence, these conversations have a rich variety of utterances that contain the essential aspects that capture how people reach an agreement in conversation. On the other hand, the conver-

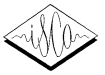
sations are on a limited domain, which should ease the problem of determining what a person's utterance is about. The limited nature should also allow us to use a smaller set of annotation tags, which just still capture the essential aspects of these conversations.

In the rest of the paper, we first describe related work in annotating what a conversation is about. We then describe the domain setup that was used for collecting the dialogues. We describe our annotation scheme in Section 4. We then report the inter-coder reliability of applying this scheme and conclude the paper.

2. Related Research

Speech act theory has been proposed to describe human behavior in conversations [7]. Speech acts indicate the role or intention of an utterance, such as informing or requesting information; speech acts also mark relations between utterances, such as an answer to a question. It is important for conversants to correctly interpret each other's speech acts in order to react accordingly. Hence, to truly understand what's going on in a conversation, dialogue researchers annotate each utterance or a group of utterances with a speech act, and analyze the patterns. A number of annotation schemes have been proposed. Carletta et al. proposed a three-level annotation scheme, and applied it on the Map Task Corpus [8]. The first level is called *move*. Each move has a speech act which is either an "Instruct", a "Question", or some other responsive tags. Moves constitute *conversational games* at the second level, which in turn constitute *transactions* at the third level. Traum and Hinkleman developed a four-level annotation scheme from the standpoint of social linguistics [9]. The basic level is utterance unit, at which utterances are given tags in terms of grounding. An initial presentation and its grounding utterances together compose a dialogue unit, which is then assigned a traditional speech act, such as "Inform" or "Request". The DAMSL scheme further advances the effort by allowing an utterance to have multiple tags, in which an utterance can be coded to have a forward function, a backward function, and a communicative function.

Speech acts, however, do not capture what participants are discussing. For instance, given that an utterance is an *Inform*, it is unclear whether it is an inform of some domain information, or of the topic that the conversants are focused on. Recently, Bates et al. proposed an annotation scheme to classify utterances based on how they contribute to a meeting [11]. Bates et al. coded a corpus of meetings with a set of *meeting acts*. They used 11 tags divided into 5 groups: administration and planning, decision making, discussion, humor, and breaks and commentary. In addition, a set of 12 diacritics was used to capture additional features, such as whether the utterance shows agreement or disagreement. Fur-



thermore, Bates et al. also tried to capture embedded discourse segments by annotating a primary and secondary tag.

When applying *meeting acts* on the corpus of meetings, however, Bates et al. found low inter-coder reliability. Bates et al. compared the annotations of three labellers with each other. The percent agreement between each pair was at most 53%, and the averaged agreement was only 47%. The problem might be that their scheme is too difficult to code with, and the domain is very complex. In comparison, the current research uses a simpler task with a simplified annotation scheme, which minimizes confusion between inter-coder, thus increasing the chance of producing higher inter-coding reliability.

3. Domain

For our work, we used a corpus of dialogues in which two people play a computer-mediated card game. In the game, the two players work together to assemble a poker hand of a full house, flush, straight, or four of a kind. Each player has three cards in their hand, which the other cannot see. Players take turns drawing an extra card and then discarding one until they find a poker hand, for which the players earn 50 points. To discourage players from simply rifling through the cards to look for a specific card without talking, one point is deducted for each new picked-up card, and ten points for a missed poker hand or incorrect poker hand. For more details about this corpus, cf. [6].

This corpus is interesting because it helps us to understand how a jointly agreed strategy is established and maintained. In the game, players converse to share card information, explore and establish strategies based on the combined cards in their hands, and might abandon an established strategy when a new strategy becomes a more viable option. Typically players would first communicate their cards, then explore different poker hands, agree on one to go for, and decide which card to discard based on this agreed strategy. Each time a new card is dealt, players would re-examine whether the established strategy is still optimal, and abandon it if a more viable one emerges.

During the card game, the players were interrupted to perform a short task independent of the card game. We have excluded utterances related to these short tasks from analysis.

4. Annotation Scheme

The study focuses on the agreement process between two players collaborating to complete a card game but has no visibility to cards each other is holding. To achieve the mutual goal they have to share information with each other, explore different strategies of pursuing a particular poker hand, and suggest a strategy with the highest probability that allows them to complete the card game in the shortest amount of time. A strategy is said to be established when it is agreed upon by both players. Re-exploration, suggestion and establishment of a new strategy is necessary when another strategy becomes a more viable option. Maintenance is the step in the process where players focus on keeping cards that will help them complete the game via an established strategy.

Hence, the agreement process consists of five simple steps, namely *information sharing*, *exploration*, *strategy suggestion* and *establishment*, and *maintenance*.

Information Sharing (IS): utterances that directly discuss what card(s) an individual has in their hand not biased toward a poker hand. Information sharing typically happens at the beginning of

each card game, as shown in Figure 1 from u1.1 to u1.4, where players communicated what cards they had in their hands, both the dominates and the suites. Also each time players picked up a new card, they would inform the other what it was; or even summarized the cards in their hands in case the other player forgot. A special case was that players inform the the other which card was being discarded when there is no established strategy, an example of which is u1.5 in Figure 1. We view this as indirectly notifying the other player of what cards remained in their hand. Thus it is also coded as information sharing.

u1.1	A:	alright so I have two fives, a six, and a Jack	IS
u1.2	B:	I've got a two, a seven, and a King	IS
u1.3	A:	how are your suits looking?	IS
u1.4	B:	random	IS
u1.5	A:	I'll get rid the six	IS

Figure 1: Examples of Information Sharing

Strategy Exploration (EXP): utterances that are used to explore the merits of a poker hand, to imply the probability of achieving a poker hand, or to compare possible strategies. In the example of Figure 2, player A has the nine of hearts, Queen of clubs, and King of hearts; player B has five of hearts, seven of diamonds, Queen of diamonds, and King of diamonds. From u2.3 to u2.5, the players were exploring different strategies. In u2.3, B summarized the cards in the hands of both players, implying that a full house is a possible option. Then B continued to explore the possibility of getting a flush because he had three diamonds in his hand. Player A responded in u2.5 providing information regarding B's exploration.

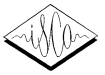
u2.1	B:	I have five, seven, Queen, King	IS
u2.2	A:	I have nine, Queen, King	IS
u2.3	B:	so we have two Queens and two Kings	EXP
u2.4	B:	I have three diamonds	EXP
u2.5	A:	no diamonds here	EXP
u2.6	B:	we'll go for Queens and Kings	SSUG
u2.7	A:	okay	EST
u2.8	B:	I am getting rid of the five	M

Figure 2: Example of Strategy Exploration and Suggestion

Strategy Suggestion (SSUG): utterances that are used to propose a certain strategy to pursue. These utterances usually start with "let's", "I think we should", "why don't we go for", etc. A strategy suggestion should be able to be followed by a "yes" or "okay" in which it is clear that the "yes" or "okay" signals that the strategy is **established (EST)**. Utterance u2.6 and u2.7 in Figure 2 show an example for SSUG and EST. After exploring a full house and a flush of diamonds, player B proposed going for a full house because it waited for only one more card, either a queen or a king, which was accepted immediately by player A. Note that in case that a strategy is not accepted, players would further explore or suggest other alternatives.

Players might implicitly suggest a strategy by stating what cards they need, as shown in Figure 3. In u3.3, player A implied that they should wait for a ten to have a straight, and player B accepted this suggestion by saying "okay" in u3.4, after which a strategy to go for a straight of "7, 8, 9, 10, J" was established.

Strategy Maintenance (M): utterances that are made to assist the completion of a strategy that has been established. These include:



u3.1	A:	I have seven, nine, and two Aces	IS
u3.2	B:	I have two, eight, and Jack	IS
u3.3	A:	we just need a ten then	SSUG
u3.4	B:	okay	EST

Figure 3: Example of Implicit Strategy Suggestion

(1) utterances stating which cards are still needed to complete an established strategy, such as “we are waiting for a King to get the straight”; (2) utterances in which a player recommends which card the other person should get rid of, such as “get rid of the seven”; (3) utterances in which a player states which card is being discarded, such as u2.8 “I am getting rid of the five” in Figure 2; and (4) utterances that indicate the completion of a poker hand, such as “we now have a straight” or “we are done”. To distinguish strategy maintenance from information sharing or strategy suggestion, a key point is that there exists an established strategy, and the intention behind the utterance is to carry out this strategy. For example, the utterance “we are waiting for a King” can be interpreted as either strategy suggestion or strategy maintenance, based on whether there is an established strategy that can be completed with a King.

5. Reliability Evaluation

5.1. Materials

We chose an excerpt of the card-game dialogue as the materials to evaluate the inter-coder reliability of the annotation scheme. This excerpt of dialogue lasted for about 318 seconds, in which the players together completed five poker games. The tool DialogueView [12] was used to organize and print out the dialogue. Annotators were given the printed-out dialogue transcripts, segmented into 258 utterances, together with the domain information such as what cards the players had in their hands, which card was discarded, what was the new card just picked up, etc. There were 92 utterances that have no forward impact on the card game, such as acknowledgement and simple repetition that signals understanding. These utterances were printed out in special color, and annotators were told not to code these utterances.

5.2. Annotators

We had five annotators applying the annotation scheme to the dialogue excerpt. Three of them are the authors of this paper, and are classified as expert annotators. We also had two other people, not involved in this project, annotate the data. Annotator 4 has extensive experience in speech act annotations, while annotator 5 has no experience. Both of them have a background in linguistics. They were first given a three-hour training session about the annotation scheme and the domain, in which we had them do an initial training excerpt, and then all five annotators reviewed a gold-standard produced by the three experts. After that, all five annotators spent one hour to code the materials independently.

5.3. Agreement between Expert Annotators

The first and third authors (annotator 1 and annotator 3) are the most familiar with the annotation scheme. Hence we first examined their inter-coder reliability. Their percent agreement is 86% (142/166), which corresponds to a Kappa of 0.80. According to [13], Kappa above 0.80 is viewed as strong agreement.

We also looked at where the differences were. Table 1 is the confusion matrix. Because most strategy establishment (EST) are acknowledgements like “okay” which were excluded for annotation, in our inter-coder reliability evaluation, we did not compare EST.

Table 1: Confusion Matrix between Annotator 1 and 3

	IS	EXP	SSUG	M	total
IS	51	9	1	1	62
EXP	2	40	0	0	42
SSUG	0	0	9	3	12
M	2	3	3	42	50
total	55	52	13	46	166

We see there are 24 differences between these two annotators. The two annotators discussed their annotation for the 24 discrepancies. In five cases, one of the annotators realized that their annotation was wrong and the other was correct. Seven cases were differences in how a response was coded. One of the annotators coded the responses with the same code as the questions, while the other coded the responses as information sharing. Hence these differences could have been easily resolved with more explicit instruction. In 9 cases, the two annotators did not agreed on what the player’s intention was. For example, one speaker said “six”, which was the last card that they needed. It was unclear if this was maintenance or information sharing, as it was unclear whether the player was communicating that they finished the hand. It is possible that listening to the audio would have resolved this ambiguity as well as others. In the last three cases, the annotators agreed on the player’s intention but not on how it fits into the annotation scheme. For example, in one utterance, a player said “and I am still keeping three diamonds”. It was unclear whether he was suggesting a strategy or merely informing the other what he was doing.

We then compared the expert annotations of the second author (annotator 2) with the other two. The inter-coder agreement between the first and second author was 80% (Kappa is 0.72), and between the second and third author was 86% (Kappa is 0.80). All three annotators agreed on the same tag 76% (126/166) of the time. There was only one utterance for which they all had a different tag. Thus, all three expert annotators seemed to apply the annotation scheme with strong agreement.

Overall, inter-coder reliability between the expert annotators is very high. A small percentage of mismatches are due to annotators trying to annotate the player’s intention, which is a challenging task to do. We believe that the inter-coding reliability can be further improved if audio recording is provided to annotators.

5.4. Agreement between Non-Experts and Experts

We now look at the agreement results between non-expert annotators and experts, as shown in Table 2. The first three rows show the percent agreement between the two non-expert annotators and the three experts individually. The forth row shows the percentage of agreeing to at least one expert. The fifth row shows the percent agreement on utterances where all three annotators have the same annotations. The sixth row show the percent agreement with the majority annotations of the experts.

Generally, agreement between non-expert annotators and experts is lower than agreement between experts. This is not surpris-

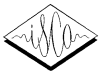


Table 2: Agreement between Non-Expert and Expert Annotators

	Annotator 4	Annotator 5
Annotator 1	72%	60%
Annotator 2	80%	60%
Annotator 3	75%	62%
any expert	85%	71%
all experts	85%	66%
majority	76%	61%

ing because (1) the non-experts might have a wrong interpretation of the players' intentions (since they were not familiar with the domain); or (2) the annotation scheme is not clear.

Annotator 4 had higher agreement with the experts' than annotator 5. This could be due to the experience of annotator 4 in tracking down the players' deep intention, such as indirect speech acts. We found that nearly half of the disagreement between annotator 5 and the experts is that the experts coded an utterance as strategy exploration while annotator 5 coded it as information sharing. For example, the surface form of utterance u2.4 in Figure 2 was to inform the other player of the cards in their hand. However, because the three diamonds that the player had was leading to a flush, and because the player only summarized a subset of his cards in hand, our experts would view it as the player exploring the possibility of a flush rather than just sharing information. This type of disagreement, in fact, is evidence that annotator 5 failed to recognize the true intention of the players.

Let's now focus on the 126 utterances where all three experts agree on the same tag (the fifth row in Table 2. Annotator 4 had the same tag 85% (108/126) of the time. Of all 18 cases of disagreement, ten were that our experts coded it as strategy maintenance while annotator 4 coded it as something else, which accounts for 56% of the disagreement. Annotator 5 agreed with the experts' 66% (83/126) of the time. Of all 43 cases of disagreement, 15 were the same type of disagreement, which accounts for 35% of the disagreement (note that 47% of the disagreement was that experts coded an utterance as strategy exploration while annotator 5 coded it as information sharing, as explained above). This shows that the two non-expert annotators were confused about the concept of strategy maintenance, although the experts had a good understanding of it. This means that we need to further clarify the concept of strategy maintenance in our annotation scheme.

6. Conclusion

In this paper, we describe an annotation scheme of coding how people reach an agreement in conversation. We have demonstrated that this annotation scheme has good inter-coder reliability in a simple yet rich domain. The annotation scheme has a concise label set, which serves as an initial attempt for understanding agreement process in human conversation. With this experience, the future work is to apply this annotation scheme in a more diverse domain, such as multi-party meetings.

7. Acknowledgements

This work was funded by the National Science Foundation under IIS-0326496.

8. References

- [1] Lance A. Ramshaw, "A three-level model for plan exploration," in *Proceedings of 29th ACL*, Berkeley CA, 1991, pp. 36–46.
- [2] Matthew Purver, Patrick Ehlen, and John Niekrasz, "Detecting action items in multi-party meetings: Annotation and initial experiments," in *The 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006)*, Washington, DC, 2006.
- [3] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act corpus," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, 2004.
- [4] Susan E. Strayer, Peter A. Heeman, and Fan Yang, "Reconciling control and discourse structure," in *Current and New Directions in Discourse and Dialogue*, J. Van Kuppevelt and R. W. Smith, Eds., chapter 14, pp. 305–323. Kluwer Academic Publishers, 2003.
- [5] 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 the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.
- [6] Peter A. Heeman, Fan Yang, Andrew L. Kun, and Alexander Shyrovkov, "Conventions in human-human multithreaded dialogues: A preliminary study," in *Proceedings of Intelligent User Interface (short paper session)*, San Diego CA, 2005, pp. 293–295.
- [7] J. R. Searle, *Speech acts: An essay in the philosophy of language*, Cambridge University Press, Cambridge, 1969.
- [8] Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, no. 1, pp. 13–31, 1997.
- [9] David R. Traum and Elizabeth A. Hinkelman, "Conversation acts in task-oriented spoken dialogue," *Computational Intelligence*, vol. 8, no. 3, pp. 575–599, 1992, Special Issue: Computational Approaches to Non-Literal Language.
- [10] Mark G. Core and James F. Allen, "Coding dialogues with the DAMSL annotation scheme," in *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, 1997, pp. 28–35.
- [11] Rebecca Bates, Patrick Menning, Elizabeth Willingham, and Chad Kuyper, "Meeting acts: A labeling system for group interaction in meetings," in *Proceedings of EUROSPEECH*, 2005.
- [12] Fan Yang, Peter A. Heeman, Kristy Hollingshead, and Susan E. Strayer, "Dialogueview: Annotating dialogues in multiple views with abstraction," *Natural Language Engineering*, To appear in 2007.
- [13] Jean Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1997.