Coding for Mutuality Strategies in Dialogue

Brian Hansen and David G. Novick
Oregon Graduate Institute of Science & Technology
20000 NW Walker Road, Beaverton, Oregon 97006 USA
(503) 690-1121
Email: brianh@cse.ogi.edu

ABSTRACT

Mutuality strategies provide a way of predicting and producing utterances that best convey the degree to which conversants have mutual understanding. We formulated the notion of a mutuality strategy model in an effort to account for the observed behaviors of human conversants engaged in task-oriented face-to-face interaction. While we have proposed to realize this model computationally and evaluate it both as a simulation of human behavior by computational agents and as a computer application, it is first necessary to establish that mutuality strategies can be identified with a reasonable level of reliability. In this paper we report the results of a study designed to determine the degree to which transcribed utterances can be coded as conveying different levels of acceptance to prior utterances.

Empirical analysis of dialogue typically has involved measuring either overall task performance or within-utterance characteristics such as disfluencies. For spoken-language systems, more useful characteristics for improving interaction would involve empirical measures that inherently relate conversants' contributions to each other. This paper looks at the feasibility of applying an extension of Clark and Shaefer's (1989) acceptance levels for conversational contributions as an empirical method of dialogue analysis. Our work is motivated by research into human-like strategies by which computers could maintain coherence. We show the application of acceptance levels as an empirical measure in a dialogue corpus and discuss the advantages and disadvantages of this approach.

For systems to interact coherently with people, they will have to use systematic approaches to achieving mutual understanding, or *mutuality.* Computer systems risk confusing their users if they jump among different methods used to achieve mutuality, especially for language-based interfaces. In order to understand what a user is saying, the system must have an expectation as to the user’s means for achieving mutuality. Likewise, to avoid confusing the user, the system must use means expected by the user under the circumstances. If our computer systems are to be adapted to the needs of humans, the answer of how to achieve mutuality in human-computer collaborative tasks must be grounded in empirical methods.

We have claimed that conversants engage in a variety of strategies for achieving mutual understanding that reflect the conversants' acceptable levels of uncertainty for each level of granularity within a task domain [4]. We proposed a model to account for patterns of interaction in dialogue, based upon conversants acting in the face of incomplete information and uncertain belief. A model, patterned after Clark and Shaefer’s [3] levels of acceptance, can be used in both the prediction and the production of utterances between collaborating conversants.

Clark and Schaefer explained conversational structure in terms of a control theory in which conversation is carried out by means of *contributions*. A contribution is made up of a *presentation* (on the part of the speaker) and an *acceptance* of the presentation (from the listener). Within their theory, a single utterance acts both as an acceptance of a previous utterance and as a presentation of new information. Acceptances are defined as being at one of five levels: continued attention, next relevant contribution, acknowledgement, demonstration, or display.

In earlier research [5], we noted that the levels of acceptance that human conversants used in the course of a human-human task-based dialogue could be closely predicted by a model that extended Clark and Shaefer’s level-of-acceptance framework. We call this the mutuality strategy model (MSM). As part of our current research we propose to realize the MSM computationally and to establish both its replicative validity via simulation [5,7,8] and its applicative viability by incorporating it into an application [1]. A prerequisite for establishing the validity of the MSM is to test whether it can be reliably coded by human transcribers. A high degree of inter-rater reliability in MSM coding could also indicate that some or all of the coding process could be automated.

THE MUTUALITY STRATEGY MODEL

One problem with the level-of-acceptance model is its generality: it doesn’t make specific predictions as to which levels of acceptance conversants should produce or predict. A good starting point for reasoning about what levels of acceptance conversants might employ in order to achieve their goals with the least collaborative effort is Clark and Brennan’s [2] description of the costs associated with different kinds of interactive behavior (repair, turn-taking, and so on).

In an earlier study of human-human interaction in a simple task-oriented domain [6], we gave subjects the assignment of collaborating in what we called the “letter sequence” task: a task designed to mimic the structure of more complex collaborative tasks in which participants have incomplete knowledge. Each pair of subjects was given cards containing a sequence of 16 letters and blanks such that they could reconstruct the entire sequence only by pooling their knowledge (see Figure 1). They were asked to put the cards out of sight and to work together to reconstruct the complete sequence from memory. The task was difficult enough that conversational breakdowns were not uncommon.

Subject A:	O	_	O	S	E	_	A	G	F	H	C	W	_	E	_	_	Z
Subject B:	_	S	O	S	_	X	_	G	F	H	_	_	L	E	B	Y	_

Figure 1. Sample letter sequences

We found that the subjects interacted differently based upon their level of certainty associated with the different levels of granularity inherent in the task. From the perspective of a single conversant, the domain of the task is naturally broken down into three levels of granularity: the single letter, the sub-sequence of letters bounded by blanks, and the entire sequence. Earlier simulations [5] centered on the interaction pattern we call the “sub-sequence hypothesis” (SSH) in which, for example, a conversant retained the conversational turn as long as she or he had positive information to convey and ceded the turn upon reaching a blank.

Although the SSH successfully portrayed the pattern of interaction that occurred in several of the experiments, in many instances other patterns occurred. Conversants sometimes got stuck and entered into lengthy repair sub-dialogues, or nodded and verbally acknowledged individual letters, echoed the other’s utterances, or sometimes even chanted along. The SSH did not predict these behaviors, nor did it account for the behavior of conversants in the face of uncertainty.

Applying acceptance levels to each level of granularity of the domain provides a way of understanding how conversants might attempt to perform the letter-sequence task, and forms the basis of the notion of mutuality strategies. By associating a level of acceptance with each level of granularity of the domain, one can describe the overall mutuality strategy of a conversant. A mutuality strategy for the letter sequence task would be a 3-tuple consisting of the acceptance level to be used for each granularity level of the domain: the letter, sub-sequence and full sequence levels.

METHODOLOGY

Before realizing the MSM computationally, we performed a pilot study to determine whether the mutuality strategies that conversants used can be coded from a transcribed dialogue with a reasonably high level of inter-rater reliability. The coders were graduate school faculty, staff, and students (five altogether) who had a basic familiarity with the notion of levels of acceptance. As a practical matter, the conversants’ changes of attention and attentional state (whether they were looking toward or away from each other) were already coded in the transcripts. The coders, therefore, needed to code only verbal events.The scope of this study is limited to the question of the extent of agreement on the verbal levels of acceptance used by the conversants.

The process of coding is essentially dividable into two parts: coding the level of acceptance, and coding the level of granularity of the task. For the purposes of this coding effort, we looked only for the acceptances at the lowest domain level. Since the sub-sequences of each conversant are known, the sub-sequence acceptances can be derived. There are two sets of codings, one for each speaker.

Although there are many possible ways of performing the letter sequence task, subjects invariably worked by making one or more passes over their sequences until they felt they had achieved their joint goal. After each pass, the conversants typically went through what we have called a diagnostic phase in which they discussed where things had gone wrong and how they should proceed. These diagnoses were not coded since the MSM makes no specific predictions as to their contents.

For the purposes of this study, the five levels of acceptance defined by Clark and Shaefer were extended to include explicit disacceptances (conversants saying “No”,”Wait”,”That’s not right”, or “What?” for example). While most of the acceptance levels are straightforward, it is important to realize that a next-relevant-contribution can take the form of both statements and questions. For example, both “my next letter is ‘J’” and “What did you have next?” are coded as NRC’s. The extended levels of acceptance are shown, with descriptions, in Table 1.

Table 1: Levels of acceptance

Code	Acceptance level	Description
DIS	Dis-acceptance	correction or request for correction
CAT	Continued attention	the state of looking toward
NRC	Next relevant contribution	the next letter, sub-sequence, diagnosis, etc.
ACK	Acknowledgment	“okay”, “alright”, etc.
DEM	Demonstration	following an explicit command
DSP	Display	echoing or paraphrasing other’s utterances

After a short training session, the coders worked independently at coding the acceptances from a transcript. A portion of this transcript is included in Appendix A.

RESULTS

For the purposes of this study we chose to look at the extent to which all coders agreed on the acceptance level of each relevant utterance. Of the 62 utterances within the scope of the study, 58 were coded identically by all five coders, producing an inter-rater reliability of 93%.

Although we obtained a fairly high level of agreement, instances where coders’ judgments differed illustrated possible problems with the level of acceptance model. These problems are discussed below.

DISCUSSION

The most significant divergence in coding was in the distinction between the DSP and NRC levels of acceptance. In the case where one conversant echoes the utterance of the other (see Appendix A, events 80, 89 and 91 for examples), one coder pointed out that the conversant who echoed may not have intended to do so; they may have not realized that the other would continue to speak and so they intended to make the next relevant contribution. One implication of this line of reasoning would be to require that there be two levels of acceptance associated with each conversant: the intended and the achieved. If the echoed conversant had taken the echo to be the next relevant contribution following their utterance, he or she would eventually be forced to conclude that there was a divergence of beliefs about the sequence, and that efforts to repair might be required. Such a divergence would constitute a mismatch of mutuality strategies. Alternately they may have deduced that the echo was inadvertent and decided to treat it as an intended echo.The fact that in these cases there was no effort at repair is good but not compelling evidence that echoes of this kind should be treated as instances of the “display” level of acceptance.

One way of distinguishing between real and inadvertent echoing is to compare the time at which the utterances took place. If the utterances were made simultaneously, then they are both considered to be NRC’s. After enough time has passed that the echoer must have known that the other did indeed keep speaking, the echo must be considered as a DSP. This distinction introduces a temporal dimension into the level of acceptance model.

The other divergence in coding occurred at the beginning of a pass over the sequence. Conversant B (whose first letter was a blank) said “You start.” whereupon conversant A said the first letter of the sequence (“O”). In this instance, the saying of the letter “O” was coded as a demonstration by some coders and as an NRC by the others. Although clearly a demonstration (by definition), it also functions as a next relevant contribution. Either an utterance may have more than one acceptance level, or we need to expand our notion of what constitutes a demonstration.

Finally, one source of possible divergence was avoided entirely. Since coders were asked to make judgments only on verbal behavior, gestures of the face and hands as well as the attentional state of the conversants were not included in the study. If they had been included we may have had a significantly lower level of inter-rater reliability since it is possible to produce many acceptance levels simultaneously by the use of different “channels.” We have not yet resolved the way in which multiple, simultaneous, and possibly conflicting levels of acceptance are to be interpreted.

CONCLUSION

Although we have established a reasonably high degree of inter-rater reliability in identifying the levels of acceptance produced by conversants in the course of a face-to-face task-oriented dialogue, those cases where we disagreed revealed some troubling aspects of the level-of-acceptance model. These difficulties may require a recasting of the model in order to make it a useful construct for incorporation into the MSM.

The establishment of a clear and coherent means of identifying the level of acceptance of utterances is crucial to our research program since we plan to establish the validity of the MSM by showing that conversations between two computational agents can replicate those between two humans. Since there is such a high degree of variability in human behavior, it is generally not possible to predict the course of a dialogue exactly. We propose to measure the degree to which a simulated conversation matches its experimental counterpart by treating the production and prediction of utterances as a problem of search. We plan to measure the “distance” between the recorded and simulated conversations. The distance between predicted and actual levels of acceptance in dialogue constitutes a key metric by which to evaluate the MSM.

The MSM has the potential to advance our understanding of interaction in several important ways. Notions of turn-taking and initiative emerge from the working of the model. It is, in addition, independent of domain knowledge, modality, and dialogue context, making it a useful step toward interaction-centered system design.

REFERENCES

[1] Blandford, A., An agent-theoretic approach to computer participation in dialogue. *International Journal of Man-Machine Studies*, Vol. 39(6) (1993), pp. 965-998.

[2] Clark, H. & Brennan, S., Grounding in communication, *Shared Cognition: Thinking as Social Practice*, APA Books (1991).

[3] Clark, H. & Shaefer, E., Contributing to discourse, *Cognitive Science*, Vol. 13 (1989), pp. 259-294.

[4] Hansen, B., Mutuality strategies in dialogue, Ph.D. Dissertation proposal (1994), Oregon Graduate Institute of Science & Technology, Department of Computer Science & Engineering.

[5] Novick, D., Control of mixed-initiative discourse through meta-locutionary acts: a computational model, Ph.D. Dissertation (1988), University of Oregon. Reprinted as Technical Report No. CIS-TR-88-18 (1988), University of Oregon.

[6] Novick, D., Hansen, B. & Lander, T., *Letter Sequence Dialogues*, Technical Report No. CS/E 94-007 (1994), Oregon Graduate Institute of Science & Technology, Department of Computer Science & Engineering.

[7] Power, R., The Organization of purposeful dialogues, *Linguistics*, Vol. 17 (1979), pp. 107-152.

[8] Traum, D. & Hinkleman, E., Conversation acts in task-oriented spoken dialogue, *Computational Intelligence*, Vol. 8(3) (1992).

APPENDIX A: A portion of a sample transcript

This transcript depicts the movements of the face, body and eyes as well as the verbal behavior of two laboratory subjects carrying out the letter sequence task. The sequences are the same as those shown in Figure 1. In transcribing we were concerned with capturing synchrony and sequence of events occurring in multiple channels. The passage of time is shown as a progression of events going from the top of the transcript to the bottom. Simultaneous and overlapping events are depicted as occurring at the same or overlapping event numbers. Note that the left conversant's verbal behavior at event 84 was the making of a sound that may have been the beginning of saying the letter "F".

EVENT #	SUBJECT A			SUBJECT B		
	FACE/ BODY	EYES	VERBAL	VERBAL	EYES	FACE/ BODY
75			O			
76		to				
77					away	
78				S		
79				O		
80			O			
81			S			
82			E			
83		away		X		
84			<efff...>			
85					to	
86			A			
87					away	
88			G			
89				g(ee)		
90		to	F			
91				F		
92			H			nods
93			C			nods
94			W			
95				L		
96				E		
97				B		
98				Y	to	
99			Z			
100	nod	away			away	nods
101				yeh...		