Indicators of Conversational Interactivity in COMPS Problem-Solving Dialogues

Michael Glass[1], Jung Hee Kim[2], Kelvin Bryant[3], and Melissa Desjarlais[4]

[1] Valparaiso University, Michael.Glass@valpo.edu
[2] North Carolina A&T State University, jungkim@ncat.edu
[3] North Carolina A&T State University, ksbryant@ncat.edu
[4] Valparaiso University, Melissa.Desjarlais@valpo.edu

Abstract. This paper reports on experiments in measuring the general level of conversational interactivity in COMPS problem-solving dialogues. COMPS is a web-delivered computer-mediated problem solving chat environment for student collaborative exploratory learning. Toward the goal of computer monitoring of conversation quality, we have analyzed dialogue turns for Initiate and Respond dialogue moves as prescribed by Conversation Analysis theory. We propose that there are shallow measures of overall interactivity that correlate with how much the students are responding. This paper reports on experiments in measuring the general level of conversational interactivity in COMPS problem-solving dialogues by attempting to classify each turn as Initiate and Respond and by attempting to predict the percentage of Respond turns.

1 Introduction

The goal of the COMPS project is to provide a computer-aided instrument for collaborative learning of concepts through problem-solving dialogue [1]. The students mainly engage in typed-chat, though for some problems COMPS has specific problem-related affordances for the students to manipulate. COMPS shows the instructor the conversations in real time, permitting the instructor to intervene. An unusual feature of the COMPS online chat environment is that students type simultaneously. They can see each other's comments as they are typed in real time. This adds an interactive dimension that even spoken language does not support, since students chat simultaneously without interrupting each other.

Toward the goal of having the computer estimate in real time the degree to which students are engaged in group problem-solving activity, this project is trying to identify whether the student dialogue turns show evidence that they are responding to each other. We annotated a set of dialogues as Initiate (I) or Respond (R), exchange structure categories from Conversation Analysis [2]. In this paper we discuss the issue of building decision tree classifiers to recognize whether each turn represents an I or R dialogue move. We also endeavor to measure the percentage of R turns as general measure of interaction, without tagging individual I and R dialogue turns.

2 Background

The learning task. The data for this study come from a second semester Java programming class at North Carolina A&T. In groups of three, students solved problems in understanding a Java Swing graphical user interface. The problems were presented to the students on paper, accompanied by a picture of the GUI with its components numbered. The nature of the task was to understand and articulate the Java software structure that necessarily lay behind the interface they were seeing. The student skills that are the focus of this project are oriented toward relational understanding. This orientation guides the exercise protocol: the students were instructed to come to shared agreement, then obtain the actual answer, then reconcile their understanding with the given answers.

Conversation Analysis and Transactivity. The linguistic discipline of Conversation Analysis analyzes the exchange structure of a dialogue, dividing up the turns into three types: initiate (I), respond (R), and sometimes followup (F). These basic structural units of dialogue are the workhorse for analyzing phenomena such as turn-taking (how people arbitrate who will speak next), social loafing (who is not participating, or being lazy), and power relationships [2]. Classifying utterances as I/R/F is potentially useful for these kinds of analyses. Transactivity is a construct used by educational dialogue analysis that closely corresponds with what we are trying to detect. A dialogue move is transactive if a) it responds to another dialogue move, and b) contributes to knowledge building [3]. Classifying whether one dialogue turn responds to another is thus poentially one aspect of finding the transactive turns that are used in analysis of group cognition.

In Weinberger and Fischer's four dimensional framework for group cognition [3] transactivity is the social mode dimension: it categorizes in what ways interpersonal processes are at work in the construction of the answer. Though there are different categories of transactive contribution (from "eliciting" to "conflict-oriented consensus building") we hypothesize that simply measuring whether a turn is transactive or not could be useful for the purpose of assessing whether group cognition is happening.

Adjustment to I/R Analysis. When there are only two people, conversation Analysis practice is to annotate Initiate/Respond dyadic exchange segments (sometimes with Followup turns). In a multi-party conversation this structure breaks down. In our conversations a single statement might elicit several responses from different participants. Persons C (turn $i + 2$) and B ($i + 1$) can both respond to A (turn i). Also C can respond to B, who responded A. We thus do not try to mark dyads, simply allowing a turn to respond to the most recent earlier turns. We eliminated the F (follow up) category, what would be tagged as an F turn is the next R turn in a chain. We note that analysis of transactivity similarly admits of threaded contributions.

3 Experiment

Data, Annotation, and Features. We conducted 17 COMPS problem-solving dialogues over two semesters with the Java Swing problem. Statistics on the dialogues are in Table 1. Three of these dialogues were extensively annotated by hand by two student annotators, according to our own variety of social contribution tags (similar in idea to transitivity) that includes an I/R dimension. The remaining 14 dialogues were manually annotated with simple I/R tags.

Treating each dialogue turn as one case, we automatically tagged the following features in the transcripts. We also produced a set of cases that contained inter-turn timing features, discussed below.

Discourse markers. A binary feature identifying a discourse marker word or phrase at the start of the turn. We used an expansive lexicon of about 90 discourse markers [4], enhanced by recognizing some variant spellings via regular expressions (e.g. soooo...).

Problem domain vocabulary. A binary feature signaling the presence of problem domain words for the Java Swing GUI problem.

Task-related deixis. This problem contained multiple-choice answer letters that students frequently referred to. The small-integer labels on the different components of the Java Swing screenshot also frequently occur. Two binary features signal the presence of these deictic references, which are related to completing the task but are not part of the domain under discussion.

Overlapped typing. We detected the number of keystrokes in the turn that overlapped in time with the typing of other turns.

Emoticons. A binary feature. People put emoticons into their chat dialogue because they express affective state, meaning that interpersonal interaction is likely happening. Though rare in this corpus, they are much more prevalent in other COMPS exercises using a different student population.

Pronouns. Binary feature indicating the presence of "you," "we," or "us" in the first 7 words, which could be indicative of interactive discourse.

Question marks. Binary feature. In our coding scheme, question turns can be either responding or initiating. Asking for an explanation for somebody's mooted answer is an example of a question marked as R. Almost any question (except possibly a rhetorical one) is a sign of students engaging with each other.

Turn length. Very short turns are often acknowledgments of other turns.

Inter-turn timing. We produced a set of cases containing the delta times between successive dialogue turns. In a three-participant conversation, computing time differences A vs. B and A vs. C doubles the number of cases. One set of cases contains the delta-times for A vs. B's most recently ended turn, the second set is identical except for delta-times A vs. C. There are four delta-times in each case: A start- and end-typing times against B (or C) start- and end-typing. An issue with generating delta-time cases occurs when A's turn is marked R. It could be a response to B or C or both. Separating A's single dialogue turn into several delta-time cases means re-annotating the R turns.

Statistics from Tagged Data Analysis. Table 3 shows the prevalence of each of the features noted above. Except for emoticons, all are attested with enough frequency that they can't be excluded from consideration. The class variable is I/R. The $\frac{R}{(I+R)}$ fractions of I and R turns for the 17 dialogues are summarized in Table 2 statistics of per-session interactivity.

Classifier Training. We trained Weka J48 decision tree classifiers to identify initiate and respond dialogue turns. We chose decision trees for these first tests because many of the features are binary, e.g. presence or absence of discourse markers. We used 10-fold cross validation in all our tests. In these experiments each training case represented one dialogue turn. We tried training using only the first 3 extensively annotated transcripts and also using all 17, about 1800 cases. We also trained decision trees using the chat delta-time records generated from the first 3 transcripts, about 300 dialogue turns. As discussed above, one turn from participant A produced two training cases for delta-times A vs. B and A vs. C.

To fit the interactivity index $\frac{R}{(I+R)}$ we used Weka's M5 multiple linear regression, utilizing all the features in Table 3. Each case was one session, where binary features were turned into numerical features by counting the number of turns in that session containing the feature.

17	Sessions
1827	Dialogue turns
107	Turns per session
52	Median duration (min)
26–67	Shortest, longest (min)

Table 1. Session Statistics

1790	Turns marked I or R
0.65	Mean of all turns
0.64	Mode of 17 sessions
0.49	Minimum session
0.72	Maximum session

Table 2. Interactivity R/(I+R)

Discourse Markers	10%
Problem domain words	20%
Overlapped typed turns	47%
Task-related deixis	30%
Emoticons	1%
Question marks	14%
Pronouns	16%

Table 3. Prevalence of Features (% of Turns)

a	b	c	← `classified as`
259	384	0	`a = I`
175	1005	0	`b = R`
0	3	0	`c = N`

Table 4. Confusion Matrix for I/R Decision Tree

4 Results and Discussion

Using the non-timing features of all 17 transcripts to build an I/R classifier, a typical 10-fold cross-validated result was κ agreement of 0.26 with human raters

and F scores of 0.44 on identifying the two class labels, the confusion matrix is shown in Table 4. κ of about 0.4 and F score of about 0.6 – 0.7 was achieved among the 3 transcripts using the delta-time cases, when care was taken to remove duplicate records and identify which turn is responding to which other participant. The best decision trees using delta-time features were not startling. If A started typing 61 or more seconds after B ended, A was most likely not responding to B. If A's statement was long it was a little more likely a response to B. Without the delta-time feature, the most determinative features were: a) Absence of question marks tends to predict R, as do very short questions of up to 2 words; b) question marks on longer sentences predict I; c) deictic references to Swing components predict I. The cross-validated linear regression estimators of $\frac{R}{(I+R)}$ were not very predictive. The correlation coefficient between the data and the generated line was 0.17.

One possible explanation for our results is inconsistent hand-annotation of the I/R dimension. We will check that and possibly re-annotate. We will also experiment with the multi-turn derived features that other researchers have used, see below. The time difference feature is domain-independent and easy to reliably mechanically generate. However generating time difference records against every other participant in the conversation proved to have a methodological problem: doubling cases biases the class labels, causing the classifier to be biased in predicting the doubled cases. We are experimenting with single-valued inter-chat delta time, measured against the most proximate chat message. Although smileys are not prevalent, they do show up in the decision trees and regression equations. This leads us to expect they may be more useful predictors when they are more prevalent. Furthermore, detecting other expressions of affect could prove fruitful.

A source of difficulty is the typing overlap. We see responses being typed before the original dialogue turn has finished, meaning a response sometimes *precedes* its target in the transcript. We see students sometimes neglect to press enter, which works because everybody can see the turn anyway. This effectively concatenates multiple dialogue turns into one chat message with pauses in the middle, which we will try to detect and split.

An indication that these results could possibly be improved comes from similar experiments from other researchers. κ between 0.5 and 0.6 was achieved classifying transitivity in online chat discussions [5] and transcribed classroom discussion [6]. In both studies the class variable was presence or absence of transactivity. They were able to boost κ agreement to 0.7 using several stages of classification. In addition to some of the features we mentioned above, these researchers derived features so that one case (one annotated dialogue turn) included contrasts against previous turns. These derived features were: 1) LSA (latent semantic analysis) comparisons of the words in the current turn to the words in the previous turn (usually another speaker), two turns back, and three turns back; 2) type of speaker (student or teacher), type of speaker for previous turn, whether the speaker is the same person as for the previous utterance; 3) change of topic: whether the topic has shifted in the previous utterance. How-

ever in some COMPS tasks the concepts and objectives being reasoned about are often not situated within the conversation. For the Swing GUI problem students instead refer to multiple-choice answers by letters and to numbered items on screenshot. This may argue against the success of LSA comparisons.

5 Conclusion

We have not yet produced a reliable estimator for a conversation's relative fraction of Conversation Analysis Initiate and Respond labels, nor have we produced a reliable individual turn classifier for same. Priority tasks are 1) to verify our training data class labels with a more rigorous annotation procedure, and 2) to refine and find more shallow features that should correlate with students engaging in transactive dialogue.

6 Acknowledgments

Thank you to our hard-working students at North Carolina A&T and Valparaiso Universities. This work is supported by the Lockheed Martin Corporation under the program of Computer Science Undergraduate Researchers to North Carolina A&T State University and by the National Science Foundation under awards 0633953 to North Carolina A&T State University and 0851721 to Valparaiso University.

References

1. Desjarlais, M., Kim, J.H., Glass, M.: COMPS Computer Mediated Problem Solving: A First Look. *Proceedings of the Midwest AI and Cognitive Science Conference* (MAICS 2012), Cincinnati. 2012.
2. Stubbs, Michael.: *Discourse Analysis*, U. of Chicago Press. (1983)
3. Weinberger, A., Fischer, F. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning, *Computers & Education*, 46(1),71–95. (2006)
4. Alemany, L.A., Masalles, I.C., Cirera, L.P.: Representing Discourse for Automatic Text Summarization via Shallow NLP techniques. Unpublished PhD thesis, Universitat de Barcelona. (2005) Lexicon of discourse markers in Catalan, Spanish, and English. Retrieved March, 2014 from `http://russell.famaf.unc.edu.ar/~laura/shallowdisc4summ/`
5. Rosé, C., Wang, Y-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger A., Fischer, F.: Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *Int. J. of Computer-Supported Collaborative Learning*, 3(3), 237–271. (2008)
6. Ai, H., Sionti, M., Wang, Y-C., Rosé, C.P.: Finding transactive contributions in whole group classroom discussions. In Gomez, K., Lyons, L., Radinsky, J. (eds.), *Proceedings of the 9th International Conference of the Learning Sciences* (ICLS '10). International Society of the Learning Sciences, vol. 1 976–983. (2010)