Automated Feedback on Group Processes: An Experience Report

Marcela Borge
Pennsylvania State University
301C Keller Building
University Park, PA 16802
mborge@psu.edu

Carolyn P. Rosé
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
cprose@cs.cmu.edu

ABSTRACT
We report on an effort to evaluate the efficacy of automated assessment and feedback of the quality of collaborative discourse in the context of an online project based course. Results of automated assessment and impact on collaborative process are evaluated over a semester-long course.

Keywords
Collaborative learning, automated process assessment

1. INTRODUCTION
In this paper we report on an effort to evaluate the efficacy of automated assessment and feedback of group processes in the context of an online project based course. It is well known that the positive effects of collaborative learning are not guaranteed. Instead, those benefits depend upon the quality of collaborative interactions that occur during activity [1]. This is problematic since most students lack the cognitive skills necessary to engage in high quality collaborative interactions [3]. Research suggests that developing socio-metacognitive expertise, the ability to understand, monitor, and regulate collective thinking processes that occur during collaboration, can help to mitigate group dysfunction and optimize collaborative interactions [4].

We have been working on developing activity design models to inform the design of Computer Supported Collaborative Learning (CSCL) systems to support socio-metacognitive development [4]. In this paper, we describe an approach to automated, collaborative discourse assessment and a study we ran in a real educational environment. We focus on two areas of inquiry motivated by emerging research. First, (RQ1) How reliably can we automatically assess collaborative discussion quality and (RQ2) does automated assessment impact future performance differently than human generated feedback?

2. METHODS
2.1 Study Context
The study took place during a 16-week, introductory, undergraduate, online course on information sciences and technology. Fourty-one online students participated in the study, each belonging to one of 14 groups. As part of the course, students were required to read a chapter from the textbook or supplementary materials each week. Students were assigned to teams within the first four weeks of the semester. Then, in weeks five, seven, nine, eleven, and fourteen, students participated in a synchronous discussion related to the reading materials. The discussion sessions were held in a collaborative workspace with chat capabilities called CREATE.

2.2 Research Design
Across the five time-points during which students engaged in a collaborative chat activity, we compared the effect of four different feedback conditions on the quality of collaboration at the next time point. After each of the first four discussion tasks, groups were assigned to one of four feedback conditions that determined the type of feedback they received at that time point.

The study was run as a within-subject manipulation. The four conditions included: (1) no feedback, (2) expert feedback, (3) automated feedback, and (4) best practices. Those in condition one received no feedback about the quality of their processes. Those in condition two received feedback from trained research assistant who would analyze their processes using our coding construct. Condition three received feedback based on automated assessment of processes. Condition four was given feedback based on common strengths and weaknesses of collaborative groups [4] and not based on the group's specific processes. All feedback was worded in a consistent manner such that teams would not know what condition they received.

An assessment of group processes was conducted for each discussion based on the transcripts from the chat environment that housed the activity. Team process measures at the first time point were used to identify groups' initial strengths and weaknesses. Thus, the first assessment was treated as a baseline, and each subsequent measurement, controlling for the previous assessment, was treated as a measure of the effectiveness of the form of feedback experienced after the previous discussion.

2.3 Assessment of Collaborative Discourse Quality
After each discussion session, individual students completed an evaluation of the quality information synthesis and knowledge negotiation in their group.

In the assessment rubric, there are three categories of behavior within each of the two core capacities, with each category assessed on a five-item, ordinal scale. The first core capacity, information synthesis, consists of three categories of discourse behavior: verbal participation, developing joint understanding, and joint idea building. Verbal participation examines the amount of turns of speech contributed by each member relative to the team's total turns of speech. Developing joint understanding evaluates the extent to which teams make an effort to ensure that members fully understand the ideas presented by taking time to reword, rephrase, or ask for further clarification of shared information. Joint idea building focuses on the extent to which team members elaborate on another member's contribution in

order to ensure that information introduced by any member is not ignored or accepted, without discussion.

The second core capacity, knowledge negotiation, also consists of three categories of behavior. These categories are contributing alternative ideas, quality of claims, and norms of evaluation. Contributing alternative ideas evaluates the extent to which teams present and discuss alternative perspectives, claims, or suggestions. Quality of claims focuses on evaluating the extent to which teams provide logical, fact-based evidence and rationale. Norms of evaluation focuses on evaluating the extent to which teams adhere to social norms that promote the development of psychological safety.

Twenty percent of the total data was double coded by the research assistant and another trained graduate student to determine inter-rater reliability of the instrument: $r = 0.86$; $p < 0.001$, Kappa = 0.64; $p < 0.001$. Once each item of a core capacity is rated, they are averaged to produce a single Collaborative Discussion Quality score, which is a continuous value between 0 and 5 that we use to track improvement over time in collaborative discussion processes in the analysis below.

2.4 Automated Assessment

A key component of the study is an evaluation of an automated assessment technique. The six scales that comprise the three dimensions of each of the two core competencies in the assessment rubric were automatically predicted based on distributions of automatically predicted process codes. Training data for the macro level regression model for the 6 scales was a corpus of 13 discussions (with a total of 7015 turns) that were hand coded with a process-analysis coding scheme developed as part of this work. We built on a coding scheme developed for a laboratory study [3], but modified it for use in a real-world classroom setting. Each discussion was hand coded at the turn level using the process analysis and then assessed along the 6 different dimensions. We established inter-rater reliability for this schema of Kappa= .74, indicating substantial reliability.

The automated process analysis models were trained using the LightSIDE tool bench. We extracted a feature space consisting of unigrams, bigrams, POS bigrams, and a line length feature, and used a Logistic regression classifier with L2 regularization to avoid over-fitting. In a leave-one-team-out cross-validation, we achieved an accuracy of 86% and kappa of .77. The assessment needed in order to generate feedback for the study is at the level of the six scales that rate two core competencies, with three dimensions each. We used the counts of predicted process codes per team to predict these six scales using a separate linear function trained using a simple linear regression for each scale.

We expected a drop in performance when applying a model trained in a previous experiment. In the initial week of the study, we used the model trained on the earlier data to generate the six scores per team. In subsequent weeks of the study, we retrained the simple linear regression models to predict hand coded assessment scores from data collected in the current study during the earlier weeks of the semester. The process coding that created the predictor variables for those regression equations was computed using the original trained process coding models.

3. RESULTS

At each of four time points in the course, we collected automated assessments of collaborative process in terms of the six assessment dimensions. Each time, each of three to four groups was assigned a rating on a 5-point scale for each of the six dimensions. The same assessments were also made by human raters in order to assess the quality of the automated rating. Over time, we continued to use the original turn level process models but adapted the simple linear regressions to compute the six scale measures from the counts of the turn level codes using the hand rated data collected in the second course so far. We evaluate the quality of the automated rating by computing a kappa with linear weighting between the sets of automated ratings and human ratings. At time point one, before any data from the second instance of the course was available, the automated ratings were assessed to be at random. By time point two, the weighted kappa was .19. It was better at time point three, specifically .4. And finally, at time point four, it was up to .58. Altogether ratings for 10 sessions of the second course were needed to adapt the models and achieve a weighted kappa of .58.

Given that the automated feedback generated at early time points in the course was based on poor quality assessments, an important question is how much of a negative impact these errors cause for students. We measured the effect of the experimental manipulation using a repeated measures ANCOVA for each scale assessment separately. In each case, the dependent measure was the scale assessment at a time point rated by an expert rater, the covariate being that scale assessment at the previous time point, the independent variable being the condition that generated the feedback received by the team at the previous time point, and time point as a nominal control variable. We did not observe any consistent improvement over time or significant effect of condition on any one of the six scale assessments.

4. CONCLUSIONS

In this paper we addressed important questions related to the automated assessment of collaborative discourse quality in real educational settings. Though the automated process analysis was evaluated as very reliable within the course that provided the training data, the automated assessments in the second run of the course were initially very poor and only improved after 3 weeks of data were collected to use for adapting the prediction models.

5. ACKNOWLEDGMENTS

This work was funded by NSF grant IIS-1320064.

6. REFERENCES

[1] Barron, B. (2003). When smart groups fail. *Journal of the Learning Sciences*, 12(3), 307-359.

[2] Biber, D. & Conrad, S. (2011). *Register, Genre, and Style*. Cambridge University Press.

[3] Borge, M., & Carroll, J. M. (2014). Verbal Equity, Cognitive Specialization, and Performance. In *Proceedings of the 18th International Conference on Supporting Group Work*, 215–225.

[4] Borge, M., Ong Shiou, Y., & Rosé, C. 2015. Design models to Support the Development of High Quality Collaborative Reasoning in Online Settings. In *the Proceedings of the International Conference of Computer Supported Collaborative Learning (CSCL) 2015,* Volume 2, 427-434.