Towards Automatic Assessment for Project Based Learning Groups

Gahgene GWEON, Rohit KUMAR, Soojin JUN, Carolyn P. ROSÉ
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213

Abstract. Project course instructors routinely perform their formal assessments based on impressions formed from their mostly indirect experience with the groups they oversee. Nevertheless, even with their limited vantage point, instructors trust their ability to make assessments and regulate group work. In this paper we present a 5 dimensional assessment framework based on data from an interview study in which we investigate the assessment goals that project course instructors have. We use this framework to identify specifically where instructors’ assessments about students diverge most from that of direct observers of group work. We then demonstrate that indicators extracted automatically from recorded speech from group meetings frequently correlate better with objective observer rating of students than that of the instructor.

Keywords. Automatic assessment, project based learning, group learning

Introduction

Instructors of student project groups continuously need to assess the progress of group work because groups frequently do not function in an ideal way. Because students even at the graduate level do not always possess all of the group collaboration skills that are desirable for success, typically graduate and undergraduate programs include project courses to offer students contexts in which they have the opportunity to learn these important social skills. They do this under the supervision of an instructor who acts as a group facilitator. Although the instructors’ guidance helps students in overcoming some of the troubles that occur during group work [5][7][8], instructors may have difficulty discerning when support is needed because much of the group work is done when instructors are not present. Sometimes the problems are intentionally hidden behind the well functioning part of the group. Therefore, instructors may miss crucial opportunities to offer support and may not notice trouble within the group until the problem escalates. Our goal is to enable such episodes to be detected and addressed in a timely manner by providing instructors with insight into group processes as they unfold.

In order to ensure that technical work towards providing reports to instructors to augment their view of student groups will be useful, that work should focus on the aspects of assessment that are most problematic. However, while researchers and project course instructors alike acknowledge that assessment of group processes is a hard problem, there have so far not been studies that quantified which aspects of the assessment they are doing are most problematic. Thus, the important contributions of this paper are an in depth analysis of the assessment project course instructors are...
doing, where it falls short, and progress towards a technical solution to filling in the specific gaps we have identified. Recent advances in automatic collaborative learning process analysis [9] bring the vision of developing a tool to support automatic group assessment within practical reach. That technology has proven capable of detecting important conversational events that are indicative of successful group learning in highly controlled settings over short periods of time. This paper presents important foundational explorations into how this basic technology can be adapted and applied to a larger problem in the more realistic, and less controlled setting of a college level engineering design course.

In the remainder of the paper we describe an assessment framework based on an analysis of data from an interview study with project course instructors. We then describe how we applied this assessment framework to data from a project course in order to measure the accuracy of the assessments an instructor made in that context. Next, we compare the correlation between the instructor’s assessments with that of objective observers as well as between indicators that can be extracted automatically from recorded speech and the same observer ratings. Our findings are on average that the automatically extracted indices correlate better than the instructor ratings to the objective observer’s ratings. We conclude with directions for our continued work.

1. Study 1 : Interview Study with Project Course Instructors

In order to find out what instructors aim to know about group processes, we conducted Study 1. We began with two specific goals. First, we sought to find out about the problems that instructors face as they attempt to supervise group work. Secondly, we worked to identify general assessment categories at the group and individual level. Taking a user centered approach, we interviewed our target users, namely the instructors who teach project courses. Using a grounded theory based analysis approach, we conducted an iterative coding process that resulted in the categorization of five pairs of assessment categories which were then formalized into coding definitions that could be reliably applied by objective observers to form a gold standard assessment.

1.1. Method: data collection and analysis

Interview data was collected by a team of three interviewers who ran nine focused interviews with instructors. All the instructors had taught at least three university level group project courses. These instructors, all from the same university, included two from design, two from the social sciences, and five from engineering. The interviews lasted from 30 minutes to an hour; and six were recorded and transcribed for further analysis. The interviews consisted of background questions about the course, specific descriptions of procedures used to assign grades, as well as methods for administering peer evaluations. More details about the interview study can be found in [4].

To guard against missing important details, at least two of the three researchers conducting the interviews were present for each interview. After every 2 or 3 interviews, all three interviewers got together to consolidate the identified themes based on their respective interview notes. From this iterative process, three meta assessment categories of learning, process, and product emerged. Next, to verify that the meta categories sufficiently covered all the data, the six recorded interviews were
transcribed and segmented into sentences, yielding a total of 2320 sentences. The segmented sentences were then coded for further analysis, which we refer to as “assessment category coding”. For this assessment coding stage, we selected the sentences related to what instructors wanted to know about the student groups. Next, to differentiate group process types within the three sets, two coders assigned short descriptive labels consisting of 3~5 words to the sentences identified as belonging the 3 main categories. The short labels were grouped to form 15 detailed categories. The resulting hierarchy of finer grained categories that emerged is displayed in Figure 1.

![Figure 1. Assessment Categories](image)

Three rounds of coding occurred during the assessment category coding stage. Kappa values computed over 20% of the data are indicated in parenthesis for each round. In the first round, the sentences were each assigned to one of the three meta assessment categories in order to see if they could be reliably differentiated and also to see how much data each of the meta categories covered (0.88). In this second round of assessment coding, two coders annotated the fifteen detailed categories which are under the three meta categories to verify that they can be reliably coded (0.72). Finally, in the third round of coding, two coders coded for the five pairs of detailed categories under the meta category of “process” because the meta category of “process” is of interest for the purpose of gaining insight into the group processes (0.90).

1.2. Results

Among the three coarse grained assessment categories of learning goals, process, and product, process is of greatest interest for two reasons. First, group difficulties can be revealed through interaction processes that display such things as amount of effort offered by group members or characteristics of group dynamics. On the other hand, learning goals set by instructors or the resulting group products do not show where and when the students are having difficulty while doing group work. Although in real work settings, it is mainly the success of the final product that matters, instructors regard the process to be important for the purpose of giving students the opportunity to learn. By influencing the process, instructors have the opportunity to enhance both the learning experience as well as to facilitate the accomplishment of a higher quality product. By the time the product has been produced and the learning objectives of the course have been accomplished, it is too late for the instructors to intervene. Finally, processes are also a more appropriate focus for a tool that is meant to be general across multiple
disciplines. The same group processes are relevant in teamwork within the domains of design, behavioral science or engineering. However, the learning objectives as well as the group products differ across disciplines and within the same discipline.

The importance of looking at process was evident in the data as well. Instructors mentioned assessment categories under process more often (70% of the instances) than under learning goals (15%) or product (15%). In addition, the number of more detailed assessment categories under the general heading of process (10) was higher than those mentioned under learning goals (3) or product (2) as seen in Figure 1. Given the importance and interest, we focus on the meta-category of process rather than learning goals or product in the rest of our paper.

2. Study 2: Classroom Study

Given the five pairs of assessment categories that instructors look for when evaluating group projects, the next step involves tracking those categories. Before tracking the assessment categories using machine learning technology, we first verified that the information desired by the instructors is observable and can be reliably tracked by humans. The second study addresses this issue by having two researchers observe project group meetings and manually evaluate the targeted assessment categories.

2.1. Method

The course that provided the context for our data collection effort is a graduate level engineering course that was offered in spring 2008 where the students work on one big project sponsored by a client. Four subgroups were formed in order to carry out the project. Because the class is a project oriented class, a major component of the grade assigned by the instructor is based entirely on their productivity, and this portion of the grade is explicitly indicated by the instructors, separate from the part of the grade related to the quality of the result. There were 2 instructors and 22 students in the class. Various types of data were collected in this class, including messages on discussion boards, reports, and weekly work logs from each student. However, this information is not enough to address the gap between what instructors would like to know about the groups they are overseeing, and what they actually see. In order to get a more specific picture of what information instructors are missing, we instrumented the course in order to collect extensive observational data from the groups. Specifically, we collected audio recordings of group meetings as well as video tapes of classroom activities.

In particular, two researchers sat in during weekly group meetings and evaluated the five pairs of assessment categories identified in the first stage of our work which lasted for five weeks. Group meetings were chosen as the target of our observation since the bulk of group project work is accomplished during the group meetings, although instructors are not able to attend group meetings due to time constraints. The scoring of the ten categories was conducted in the following way. For each of the ten categories, we constructed statements that described the positive student behaviors associated with each given process. For example, for the category of “interpersonal dynamics”, one of the statements is, “Is everyone's opinion taken seriously without being ignored? Is there an attitude towards valuing everyone's suggestions?”. Based on the observations made in response to these statements, scorers assigned an overall score for the “interpersonal dynamics” category with a number
The range has both negative and positive numbers so that the scorers can easily map negative behaviors to negative scores and positive behaviors to positive scores. Five numbers can also capture the difference between behaviors sufficiently as used in many grading systems (e.g., A–F). The observations were conducted once a week, and scores were assigned for each of the 4 groups for each group level category, and for each of the 22 students for each student level category.

Therefore, at the end of five weeks, a total of 20 group level scores and 110 individual level scores were assigned by each of the two researchers for the assessment categories. These scores were used to calculate a reliability measure, which would show whether the assessments can be made reliably from these observations. The reliability of the coding scheme for the assessment categories was indeed verified by calculating the correlation between the scores assigned by the two researchers, which was 0.81 for the group level categories and 0.64 for individual level categories.

2.2. Speech Processing

Using the recordings collected from each student during group meetings, we computed an estimate of activity level for each student using machine learning technology. Average activity level is an approximate measure of the amount of talk that the student contributed during group meetings. Before this could be computed from the speech, the recordings needed to be segmented, and each segment needed to be coded for the amount of speech by the associated student that the recording was of. We chose to segment the speech into 10 second intervals so that it would be reasonable to assume that for most segments, there would be at most a single dominant speaker. That allowed us to utilize a relatively simplistic approach to coding activity level for individual segments. We adopted the following 4-point scale for activity level; 0-no speech from primary speaker; 1-primary speaker only does back-channeling, where back-channeling is a way of showing a speaker that you follow and understand their contributions, often through interjections such as “I see”, “yes”, “OK”, “uh-huh”; 2-primary speaker speaks holds the floor for less than half of the 10 seconds; 3-primary speaker speaks holds the floor for more than half of the 10 seconds. We first verified that human annotators could make this judgment reliably from the audio recordings of individual segments. Using this coding scheme, the inter rater reliability evaluated for two coders over 144 segments was 0.78 Kappa. With the reliable coding scheme, a single coder then coded 1132 segments (distributed evenly across students, from meetings during Phase 1 of the course). The largest proportion of segments was coded as 0, which amounted to 47.5% of the segments. 8.5% were coded as 1, 30.5% as 2, and 13.5% as 3.

In order to apply machine learning to speech, each segment of speech must first be transformed into a set of feature-value pairs. The activity level that we are trying to predict from speech is related to “how” the words are spoken rather than the content of the words. Such structural aspects of speech are captured by speech prosody. Similarly, other speech applications such as emotion detection are also more concerned with speech prosody rather than the content [1], and thus use features similar to ours. In contrast, speech applications such as dictation software use content related features such as spectral features processed through a speech recognition system.

Therefore, the features extracted from speech for our experiment are variations of prosodic features such as pitch, power and amount of silence. A total of 39 prosodic features were extracted for each of the 10 second segments using wavesurfer [1].
extracted pitch and power contours for each of the 10 second segments. The variations
of prosodic features that we extracted using these contours are detailed in the next
paragraph. The numbers in brackets indicate number of features. All features are
computed automatically.

For the pitch related features, we calculated average, maximum, minimum and
range of F0 (4) and delta F0 (4). All F0 parameters are computed only over voiced
frames in the utterances. The voiced frames are identified while computing F0 in
Wavesurfer. Power features were computed similarly to pitch features. They include
average, maximum, minimum and range of power (4) and delta power (4). Unlike F0
related features, these features are computed considering both the voiced and unvoiced
frames. We added total power and power in voiced frames (2) as well. Also, a ten point
power contour was calculated. The ten points are computed by dividing the segment
into uniformly sized one second sub-segments and computing the average power of
each sub-segment (10). We also computed average, range and standard deviation over
the ten point power contour (3). Average, maximum, minimum, range and standard
deviation of the deltas of the ten point power contour (5) were included. A feature that
counted the number of points on the ten point power contour which were above the
average delta was used as well (1). Finally, duration related features included duration
of voicing and duration of initial silence (2). Duration of initial silence was computed
automatically using heuristics on power and F0 contour.

With the coded speech data after it had been transformed into a vector
representation, we then evaluated whether it was possible to use machine learning to
automatically assign segments of speech to one of these four categories with high
enough accuracy. Because the feature space was small and we could not rule out the
possibility that there may be interactions between features within our vector
representation, we used Weka’s SMO (Weka’s implementation of support vector
machine) learning algorithm [10]. In order to avoid the evaluation results being
inflated due to overlap in speakers between train and test sets, we adopted a cross-
validation methodology where a model was first trained on all but one student, and then
performance was evaluated over the segments of the remaining student. We did this
for each student and then averaged across students to compute the performance of
74.26% accuracy. We then validated the model by using the human coded numbers for
each student to compute an average activity level, and then made a similar computation
using predicted values from the cross-validation experiment. When we correlated the
average activity levels for each student based on human codes with those based on the
automatic codes, we achieved a correlation coefficient of 0.97, indicating that we can
achieve a reliable estimate of activity level using a machine learning model. We then
trained a model using all of the coded data, which we used subsequently to code the
data used in the correlation analysis presented later in the paper.

For the test data, two of the five meeting recordings that were submitted in phase 2
were randomly selected for each student, then averaged to yield one score per student.
Four students never turned in any recordings, so 18 students’ recordings were
segmented into 10 second segments. The length of each recording differed due to
differences in meeting lengths. The number of segments ranged between 7 minutes 30
seconds to 2 hours 19 minutes 50 seconds in length (45 to 839 ten second segments),
with an average of 47 minutes in length (282 segments). Using the speech model built
in step3 with the Weka toolkit [10], student recordings from phase 2 were assigned
amount of talk values.
2.3. Results

Two main types of evaluative data were collected and coded using the assessment framework discussed above: namely, “direct observer observation scores”, and “instructor observation scores”. In order to collect this data, we asked the instructors and observers (the researchers referred to in section 2.1) to make weekly evaluations of students in the five areas of the assessment framework that was established through the interviews. In addition, we had “instructor grade” which were formal grades assigned to the students for the class for the duration of the observation period. By comparing the assessments made by instructors and observers we can see where the instructor’s scores diverge most from those of the direct observers, and thus where they are most in need of automatic support. We were not able to compute correlations for the Group Dynamics dimension due to insufficient variability in the instructor ratings on that dimension. Note that our analysis focuses on individual level assessments since the indices we extract from the speech are from individual recordings.

Table 1: Correlation between observer ratings for each student along dimensions of Goal setting, Progress, Knowledge Sharing, Division of Labor, and the average across dimensions. Note that the Group Dynamics dimension was left out of this correlation analysis due to lack of variability in scores.

<table>
<thead>
<tr>
<th></th>
<th>Observer Goal</th>
<th>Observer Progress</th>
<th>Observer Knowledge Sharing</th>
<th>Observer Labor Division</th>
<th>Observer Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructor observation</td>
<td>.032</td>
<td>-.347</td>
<td>-.021</td>
<td>.222</td>
<td>.292</td>
</tr>
<tr>
<td>Instructor grade</td>
<td>-.032</td>
<td>-.169</td>
<td>.366</td>
<td>.211</td>
<td>.273</td>
</tr>
<tr>
<td>Activity from speech</td>
<td>.514</td>
<td>-.031</td>
<td>.309</td>
<td>.351</td>
<td>.447</td>
</tr>
</tbody>
</table>

In our analysis we first sought to determine which of the dimensions instructors have most difficulty with from their limited vantage point. We did this by computing a Pearson’s correlation between the Observer ratings for each student on each dimension with the corresponding ratings for each dimension from the instructor. These numbers are reported in the first row of Table 1. Note that due to lack of variability of Instructor scores within the Group Dynamics dimension, we were not able to compute a correlation for that dimension, and thus we have left it out of the table. What we see here is striking. Correlations are consistently below .3, sometimes essentially 0, and other times even negative! This shows that instructors views on each assessment categories do not line up very well with those who observed student group meetings directly. In addition to the individual scores instructors assigned per student related to the 5 assessment dimensions, we also computed a correlation between the instructor assigned official grades and the observer ratings for those students on each dimension. We see these results in the second row of Table 1. Again, the correlations are not too strong. Only the Knowledge Sharing dimension showed a substantial improvement in predictive power over the more specific instructor observation scores which were assessed for each dimension. Finally, we computed a correlation between the level of speech activity indicator computed from the meeting recordings with the observer ratings. These numbers are reported in the third row of Table 1. Notice that despite being a totally automatically computed indicator, this was the best performing indicator in the Goal Setting, Division of Labor, and Average categories. For Knowledge
Sharing, it was not substantially different from that of Instructor grade. The only place where it under-performs is on the Progress dimension where its correlation with observer ratings is essentially 0. These correlation scores imply the potential of automatic assessment from speech as an additional source of information that the instructors can use in evaluating students.

3. Conclusions and Current Directions

In this paper, we describe a technique for predicting amount of speech activity level in recordings of group meetings. Our results suggest that quantities such as amount of speech activity, which can be computed directly from recorded speech, are useful for making assessments of group work. The next step in our work will be to test how providing information we can extract from the recorded speech influences instructors’ behavior and assessments. Further work can also be done to identify other quantities, apart from amount of speech activity, which can be extracted from the speech that might be useful for project course instructors.

Acknowledgements

This research was funded by NSF EEC grant number 064848.

References