Identifying Student Contributions through Automated Team Summaries

Kai Presler-Marshall, Bowdoin College
Sarah Heckman, North Carolina State University
Kathryn T. Stolee, North Carolina State University

Questions? Rude Comments? k.preslermarshall@bowdoin.edu
Motivation

• Teaming is a core element in professional software engineering
• Thus, it’s essential for CS and SE programs to teach students how to work in teams
• Some students may be inclined to freeride off of the contributions of their peers, receiving a grade not commensurate to their contributions
Motivation

- Accurately identifying students’ contributions to team projects remains an open challenge.
- Thus, teaching assistants may struggle to give students consistent & actionable feedback on their contributions.
- *Can autogenerated summaries of students’ code contributions assist TAs in giving better feedback?*
Course Context

- Sophomore-level Java programming course
- Lecture & projects
- **Associated lab section**
  - Students work on labs in small teams
  - Lab grading is *mostly* automated

April 24, 2023
Selected Research Questions

- Can automated summaries of student contributions enable:
  - **RQ1**: Faster grading by TAs (short answer: no)
  - **RQ2**: More consistent grading by TAs?
  - **RQ3**: Less frustrating grading from perspective of TAs?
  - **RQ4**: Better feedback for students?
public class KaiDemoClass {
    private String aField = "demo";
    public String capitaliseAndReturn(){
        String modified = aField.toUpperCase();
        return modified;
    }
}

public class KaiDemoClass {
    private String aField = "demo";
    private String secondField = "moreDemo";
    public String capitaliseAndReturn(){
        String modified = aField.toUpperCase();
        return modified;
    }
    public Object anotherMethod(){ return null; }
}
I recruited 13 former or current CS TAs
  – 12 of 13 TAs had experience grading team-based projects
I tasked participants with:
  – Grading projects
  – Considering feedback from their peers
  – Reflecting on the experience

Study Outline

Introduction (~10m)
Part 1: Grading (~90m)
  RQ1, RQ2, RQ4
Part 2: Evaluating Feedback (~10m)
  RQ4
Reflection (~5m)
  RQ3
### Part 1

<table>
<thead>
<tr>
<th>Task Number</th>
<th>Grade Contributions After this timestamp</th>
<th>Automated Summaries</th>
<th>Repo</th>
<th>Student 'A' Credit</th>
<th>Why did you give 'A' the grade you did?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>9/14/2021 10:20 AM</td>
<td><a href="https://example.github.com/">Link</a></td>
<td><a href="https://example.github.com/">Link</a></td>
<td>5</td>
<td>They tested some valid and invalid scenarios, but didn’t do as much as their teammates.</td>
</tr>
<tr>
<td>1</td>
<td>11/16/2021 10:20 AM</td>
<td><a href="https://example.github.com/">Link</a></td>
<td><a href="https://example.github.com/">Link</a></td>
<td>10</td>
<td>They seemed to make reasonable contributions.</td>
</tr>
<tr>
<td>2</td>
<td>10/25/2021 7:10 PM</td>
<td>No automated summary available</td>
<td><a href="https://example.github.com/">Link</a></td>
<td>10</td>
<td>Reasonable contribution on both implementation + testing.</td>
</tr>
<tr>
<td>3</td>
<td>9/12/2021 12:00 AM</td>
<td>No automated summary available</td>
<td><a href="https://example.github.com/">Link</a></td>
<td>10</td>
<td>Reasonable contribution on both implementation + testing.</td>
</tr>
<tr>
<td>4</td>
<td>10/13/2021 2:00 PM</td>
<td><a href="https://example.github.com/">Link</a></td>
<td><a href="https://example.github.com/">Link</a></td>
<td>0</td>
<td>We would like to see better team contributions next lab. Please work for your team to split the tasks.</td>
</tr>
<tr>
<td>5</td>
<td>11/13/2021 12:00 AM</td>
<td><a href="https://example.github.com/">Link</a></td>
<td><a href="https://example.github.com/">Link</a></td>
<td>10</td>
<td>Reasonable contribution on both implementation + testing.</td>
</tr>
<tr>
<td>6</td>
<td>11/1/2021 10:20 AM</td>
<td><a href="https://example.github.com/">Link</a></td>
<td><a href="https://example.github.com/">Link</a></td>
<td>5</td>
<td>Good contribution to the implementation this week.</td>
</tr>
<tr>
<td>7</td>
<td>11/1/2021 12:00 AM</td>
<td>No automated summary available</td>
<td><a href="https://example.github.com/">Link</a></td>
<td>10</td>
<td>Reasonable contribution on both implementation + testing.</td>
</tr>
<tr>
<td>8</td>
<td>11/1/2021 12:10 AM</td>
<td>No automated summary available</td>
<td><a href="https://example.github.com/">Link</a></td>
<td>5</td>
<td>Good contribution to the implementation this week.</td>
</tr>
<tr>
<td>Grade 1</td>
<td>Comment 1</td>
<td>Grade 2</td>
<td>Comment 2</td>
<td>I would choose</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>Try to contribute more by coordinating with your teammates and asking what help is needed.</td>
<td>5</td>
<td>Good job on fixing code and adding tests! Next time see if you can contribute more on the implementation side of things.</td>
<td>Either (no difference)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>good job in FSM and code contribution</td>
<td>10</td>
<td>Great work both on implementation and testing.</td>
<td>Either (no difference)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Good job in implementation and testing</td>
<td>10</td>
<td>There is some implementation and testing along with the Javadocs. But it would be better if you did some more implementation.</td>
<td>Either (no difference)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>fixing checkstyle is not enough contribution</td>
<td>0</td>
<td>you need to write more tests and implementations rather than fixing typos and generating Javadocs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>It looks like all you mostly did was documentation this week. In the future please try and split the tasks up so everyone has a coding portion.</td>
<td>5</td>
<td>Great work on implementation and testing.</td>
<td>Comment 2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>Done nothing</td>
<td>10</td>
<td>Good work on making sure the runtime is updated with progress, and generating documentation. However, try to contribute more on the implementation or testing as well.</td>
<td>Comment 1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>fixing typos and capitalization is not a valid contribution</td>
<td>0</td>
<td>Excellent work on making contributions on the implementation, testing, and the documentation. Keep up the good work.</td>
<td>Comment 1</td>
<td></td>
</tr>
</tbody>
</table>
RQ2: Grading Consistency

- TAs grade projects much more consistently ($q = .021$) with contributions summaries to assist them
- However, consistency still remains a challenge even with contributions summaries ($\alpha = .609$)
All participants prefer grading with automated summaries; 11/13 strongly prefer them.

RQ3: Grading Preferences
RQ4: Feedback Quality

- TAs consider feedback from assignments graded with contributions summaries more actionable ($p = .031$)
- TAs provide more partial credit when grading with contributions summaries ($p = .018$)
  - Since we require TAs to provide feedback with partial credit, but not full credit, this shows they can improve the quantity of feedback provided
Discussion & Future Work

- Despite a small sample size & relatively primitive summaries algorithm, a lab study showed value of my contributions summary algorithm
- I am running a follow-on classroom study
  - Same experimental & control groups
  - Do students find feedback more actionable? Do they improve more over the semester?
  - Do we get the same consistency benefits with an entire semester of assignments?
Discussion & Future Work

- There is, of course, a lot more that goes into SE work than just Java code.
- How can we efficiently handle other types of (code) contributions? Can language-agnostic AST analysis help with scalability?
- Can we account for all of the other (non-code) contributions to a SE project?
- These are some of the questions I’m hoping to ponder in detail this summer.
Summary

- I designed an algorithm to summarise individual students’ contributions to team projects, and built it into a tool, AutoVCS.
- Through a quantitative lab study, I demonstrated that TAs who use these summaries grade more consistently, provide feedback that is possibly more actionable, and they prefer the grading process.