Instant Feedback for Increasing the Presence of Solutions in Peer Reviews

Huy Nguyen¹ and Wenting Xiong³ and Diane Litman¹²

¹Computer Science Department, University of Pittsburgh, PA, USA
²Learning Research & Development Center, University of Pittsburgh, PA, USA
³IBM Watson, Yorktown Heights, NY, USA
E-mail: {hvn3,dlitman}@pitt.edu, wentingxiong@gmail.com

Abstract

Educational studies have identified the presence of solutions to problems as one of the most desirable features of student peer reviews. We present the design and evaluation of a web-based peer review system that uses natural language processing to automatically evaluate and provide instant feedback regarding the presence of solutions in peer reviews. Student reviewers can then choose to either revise their reviews to address the system’s feedback, or ignore the feedback and submit their original reviews. A system deployment in multiple high school classrooms shows that our solution prediction model triggers instant feedback with high precision, and that the feedback is successful in increasing the number of peer reviews with solutions.

1 Introduction

Peer review provides learning opportunities for students in their roles as both author and reviewer, and is a promising approach for helping students improve their writing (Lundstrom and Baker, 2009). However, one limitation of peer review is that student reviewers are generally novices in their disciplines and typically inexperienced in constructing helpful textual reviews (Cho and Schunn, 2007). Research in the learning sciences has identified properties of helpful comments in textual reviews, e.g., localizing where problems occur in a paper and suggesting solutions to problems (Nelson and Schunn, 2009), or providing review justifications such as explanations of judgments (Gielen et al., 2010). Research in computer science, in turn, has used natural language processing and machine learning to build models for automatically identifying helpful review properties, including localization and solution (Xiong and Litman, 2010; Nguyen and Litman, 2013; Xiong et al., 2012; Nguyen and Litman, 2014), as well as quality and tone (Ramachandran and Gehringer, 2015). While such prediction models have been evaluated intrinsically (i.e., with respect to predicting gold-standard labels), few have actually been incorporated into working peer review systems and evaluated extrinsically (Ramachandran and Gehringer, 2013; Nguyen et al., 2014).

The SWoRD research project¹ involves different active research threads for improving the utility of an existing web-based peer review system. Our research in the SWoRD project aims at building instant feedback components for improving the quality of textual peer reviews. Our initial work focused on improving review localization (Nguyen et al., 2014). Here we focus on increasing the presence of solutions in reviews. When students submit reviews, natural language processing is used to automatically predict whether a solution is present in each peer review comment. If not enough critical comments are predicted to contain explicit solutions for how to make the paper better, students are taken from the original review interface to a new instant feedback interface which scaffolds the student in productively revising the original peer review (Figure 1).

Section 2 describes the Instant-feedback work-flow and user interface, while Section 3 describes the supporting natural language processing techniques. Section 4 then presents evaluation results

¹https://sites.google.com/site/swordlrdc/new-features
Assignment Description
The passage below is from Last Child in the Woods (2008) by Richard Louv. Read the passage carefully. Then, in a well-developed essay, analyze the rhetorical strategies Louv uses to develop his argument about the separation between people and nature. Support your analysis with specific references to the text ...

1. Thesis
Provide feedback on the quality of the author’s thesis.

Thesis. Did the author include a clear, specific thesis in his or her introduction?
1 – The author did not include a thesis in his or her introduction
...
Solution, Problem-only, or Non-criticism. Problem-only comments point out problems in the reviewed paper without providing any solutions, while Non-criticisms such as summaries or praise do not require solutions. If the review is predicted to have a S-RATIO which is the ratio of solution comments over the sum of solution and problem-only comments, less than or equal to a threshold\(^4\) (Figure 2, diamond #2), instant feedback is triggered to scaffold the student in revising the problem-only comments. Otherwise, the review is deemed acceptable and stored for later use by SWoRD Phase 3.

When instant feedback is triggered, the instant feedback interface (Figure 1, right) displays a message at the top suggesting that comments may need to be revised to include solutions, followed by buttons representing the 3 possible reviewer responses: revise the review and resubmit (left), view some predefined example comments with solutions before responding (center), or submit the review without revision (right). To call the reviewer’s attention to comments that might need revision, the interface turns text boxes around predicted problem-only comments to red (Figure 1, middle right). For these comments, the system also generates option buttons that ask reviewers to provide feedback on the prediction. We hypothesized that asking students to reason about the absence of solutions in their own comments would promote review revision. Their feedback on the system’s predictions also provides new annotated examples for future re-training of the prediction model (described in Section 3). Conversely, the interface highlights predicted solution comments in green (Figure 1, bottom right) along with displaying a thumbs-up icon. This highlighting was designed to draw reviewer attention to examples of solutions in their own comments. Finally, for reviews that are revised and resubmitted, Instant-feedback SWoRD increases the submission order and re-checks the threshold (diamond #1 in Figure 2). Unrevised reviews are instead stored for Phase 3 of SWoRD.

### 3 Comment-level Solution Prediction

To support the instant feedback interface described in the prior section, we developed a 3-way classification model for predicting a review comment’s feedback type: Solution, Problem-only, or Non-criticism. Challenges emerge from the fact that SWoRD serves a wide range of classes ranging from high school to graduate school and from STEM to language arts. Consequently, our prediction model has to process peer review comments that greatly differ in style and vocabulary. We thus focused on modeling how students suggested solutions by developing the following feature sets that abstracted over specific lexicons and paper topics:

- **Simple**: word count and order of the comment.
- **Keywords**: we semi-automatically created 10 keyword sets to model different content patterns (Xiong et al., 2010): Solution, Idea, Suggestion, Location, Connective, Positive, Negative, Summary, Error, Negation. For each set, we count the total occurrences of its keywords in the comment.
- **Location phrases**: we observed in our training data that solution content usually co-occurs with location information in comments. Thus, we extracted words and phrases that signal positional localization in comments of training data. This feature set includes hand-crafted regular expressions of location patterns (e.g., on page 5) (Xiong et al., 2010), location seed words (manually collected, e.g., page, thesis, conclusion), and location bigrams (automatically extracted given the location seeds, e.g., transition paragraph). For each location seed/phrase or expression, we count its occurrences or matches in the comment.
- **Paper content**: motivated by topic word features in (Kim et al., 2006), this feature set was designed to model how much of a paper’s content/topic was mentioned in the comment. We first extracted bigrams with TF-IDF above average in the training data, and collected unigrams that make-up these bigrams. Domain unigram feature is the number of collected unigrams in the comment. Window size feature is the length of maximal common text span between the comment and the paper (Ernst-Gerlach and Crane, 2008). Similarity feature searches for the highest similarity score between paper sentence to the comment. We extract 5 paper sentences (1 covering the common span, 2 preceding, and 2 following). For each pair of paper sentence and the comment, we apply different similarity scores (e.g., Levenshtein, cosine) to 4 abstractions of the pair (sequence of tokens, sequence of part-of-speech, sequence of nouns, sequence of verbs), and return the pair’s sum score. Feature value is the highest sum score over all pairs.

Our solution prediction model was trained with logistic regression using annotated peer review comments from two university (Computer Science, History) and a high-school class (Literature). During

\(^4\)For the deployment in Section 4, S-RATIO was tuned to 0.7 using development data from prior classes.
Table 1: Comment-level solution prediction performance. Acc: Accuracy, F1 by class label is reported – Sln: Solution, Prb: Problem-only, Non: Non-criticism.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc.</th>
<th>$\kappa$</th>
<th>F1:Sln</th>
<th>F1:Prb</th>
<th>F1:Non</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>0.50</td>
<td>0.24</td>
<td>0.40</td>
<td>0.51</td>
<td>0.57</td>
</tr>
<tr>
<td>SWoRD</td>
<td>0.62</td>
<td>0.44</td>
<td>0.55</td>
<td>0.59</td>
<td>0.72</td>
</tr>
</tbody>
</table>

4 Preliminary Evaluation

In Spring 2015, SWoRD with instant-feedback was deployed in 9 high-school Advanced Placement (AP) classes. We conducted preliminary evaluations to answer two research questions: (1) How precisely does the system predict peer review solution and trigger the instant feedback? (2) How does the instant feedback impact review revisions?

We collected peer review submissions which were intervened by Instant-feedback SWoRD (i.e., failed the S-RATIO check, Figure 2), and their immediately subsequent resubmissions (if any), then had an expert manually code the collected comments for their feedback types: solution, problem-only, non-criticism (double-coded data had inter-rater $\kappa = 0.87$).

Only intervened submissions were used to evaluate the model performance because the subsequent resubmissions were not predicted. In our deployment, 134 review submissions were intervened, containing 891 comments: 223 Solution, 340 Problem-only, and 328 Non-criticism. Table 1 shows that our deployed model outperforms a Bag-of-Words (BoW) baseline\(^5\) in predicting solution in review comments. Given the fact that the AP data was never used for training our model, the obtained performance at comment-level is promising and encourages us to improve the model with more data.

Regarding instant feedback precision, we calculated the true S-RATIO for each intervened review (using gold standard labels). Table 2 shows that given the 0.7 threshold used for this deployment, Instant-feedback SWoRD incorrectly triggered instant feedback for 24 submissions (column 3) out of 134, yielding a precision 0.82. Because Instant-feedback SWoRD does not let student reviewers know the S-RATIO threshold, students should only think that the instant feedback was incorrect when they provided solutions for all mentioned problems (true S-RATIO = 1). From this student perspective, Instant-feedback SWoRD had 16 incorrect triggers (column 4), achieving a precision 0.88.

Finally, to evaluate the impact of instant feedback on review revision, we considered the 74 subsequent resubmissions. We collected comments that were revised or newly-added to the resubmissions (no comment was deleted), and obtained 115 comments. Pairing 111 revised comments with their original versions, we observed that 73 (66%) comments were fixed from problem-only to solution, 3 (3%) from non-criticism to solution, only 1 comment was (0.9%) edited from solution to non-criticism, and none from solution to problem-only. All of the 4 newly-added comments mentioned problems and provided solutions. These results suggest that Instant-feedback SWoRD does indeed help reviewers revise their comments to include more solutions.

5 Conclusions and Future Work

This paper presented Instant-feedback SWoRD, which was designed to increase the presence of solutions in peer reviews. Evaluation results showed that Instant-feedback SWoRD achieved high performance in predicting solution in review comments and in triggering instant feedback. Moreover, for reviewers who revised their reviews after receiving instant feedback, the number of comments with solution increased. In future work, we plan to use more data from a wider range of classes to re-train the currently deployed prediction model. Also, a comprehensive comparison of our approach to studies of similar tasks would give us insight into features and algorithms for performance improvement.

\(^5\)Used 1,2,3-grams as features, and was trained with the same algorithm and data as our model.
References


