Instant Feedback for Increasing the Presence of Solutions in Peer Reviews

Huy Nguyen[1] and **Wenting Xiong**[3] and **Diane Litman**[1,2]
[1]Computer Science Department, University of Pittsburgh, PA, USA
[2]Learning Research & Development Center, University of Pittsburgh, PA, USA
[3]IBM Watson Health, Yorktown Heights, NY, USA
E-mail: {hvn3,dlitman}@pitt.edu, wxiong@us.ibm.com

Abstract

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

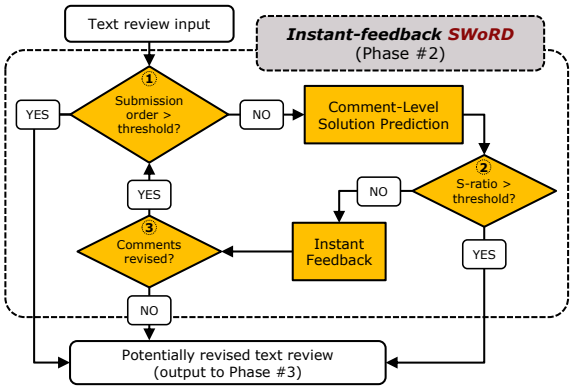

Figure 1: Architecture of Instant-feedback SWoRD.

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 project[1] 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,

[1] *https://sites.google.com/site/swordlrdc/new-features*

Proceedings of NAACL-HLT 2016 (Demonstrations), pages 6–10,
San Diego, California, June 12-17, 2016.

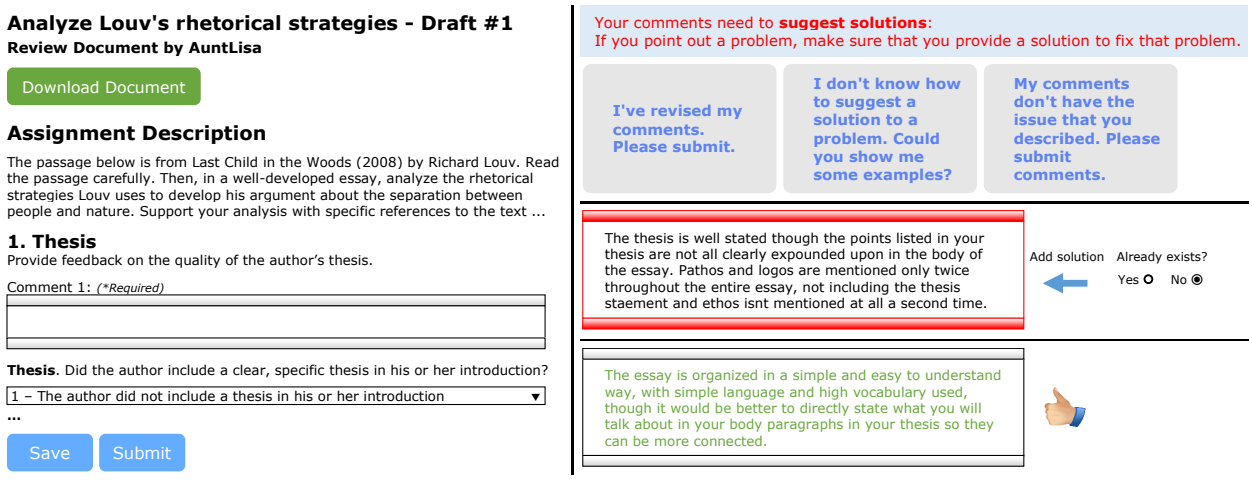

Figure 2: Screenshots of original **review interface** (left) and new **instant feedback interface** (right). For readability, the review interface shows only one comment prompt and its associated rating prompt. The instant feedback interface displays a solution feedback message and three possible reviewer reactions (top), and highlights *problem-only* (middle) and *solution* (bottom) comments.

natural language processing is used to automatically predict whether a solution is present in each peer review comment (Figure 1). 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 them in productively revising the original peer reviews (Figure 2).

Sections 2 and 3 describe the Instant-feedback workflow, and the supporting natural language processing techniques. Section 4 demonstrates the promise of our system in supporting student review revision in a recent system deployment.

2 Instant-feedback SWoRD

SWoRD[2] was developed to support web-based reciprocal peer review, especially in large classes involving writing in the disciplines where writing and revision are hard to support due to lack of resources. A typical peer review exercise in SWoRD involves three main phases: (1) student authors submit papers to SWoRD, (2) student reviewers download papers assigned to them and submit peer reviews of the papers, and (3) student authors submit paper revisions that address the peer reviews they received. To further enhance the utility of SWoRD, we have developed *Instant-feedback SWoRD*, with the goal of helping student reviewers increase the presence of solutions in the peer review comments produced during Phase 2 of the typical peer review exercise.

Figures 2 and 1 illustrate technical details of Instant-feedback SWoRD. As in the original SWoRD, student reviewers create a new review session by opening the review interface (Figure 2, left). Now, however, whenever the SUBMIT button is clicked, the "text review input" is passed to the "Submission order check" (Figure 1, diamond #1). The submission order threshold[3] specifies how many times a review will be processed for instant feedback (e.g., 0 means no instant feedback, 1 means only the original comments are analyzed, 2 means revised comments are also analyzed, etc.). If the threshold is not reached, each comment in the review is analyzed by the "Comment-level Solution Prediction Component" (see Section 3) and classified as a *Solution*, *Problem-only*, or *Non-criticism*. *Problem-only* comments point out problems without providing solutions, while *Non-criticisms* such as summaries or praise do not require solutions. To measure how many problem comments have solutions, we define S-RATIO as number of *solution* comments over the sum of *solution* and *problem-only* comments. If the predicted S-RATIO is less than or equal to a threshold[4] (Figure 1, diamond #2), instant feedback is trig-

[2]SWoRD is now licensed by Panther Learning Systems Inc. – *www.peerceptiv.com*. A free version for users willing to trial instant feedback is available at *https://sword.lrdc.pitt.edu*.

[3]The deployment in Section 4 used a threshold of 1.

[4]For the deployment in Section 4, S-RATIO was tuned to 0.7

gered to scaffold students in revising *problem-only* comments. Otherwise the review is deemed acceptable and stored for later use by Phase 3.

When instant feedback is triggered, the instant feedback interface (Figure 2, 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 2, 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 2, 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 1). 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, extending prior work (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 regular 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, e.g., '*civil*' and '*war*' in '*civil war*'.[5] *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 classes (Computer Science, History) and a high-school class (Literature). During learning, we used a cost matrix to favor instant feedback precision over recall by penalizing relevant error types. We thought it would be better to miss some feedback opportunities than to incorrectly trigger instant feedback (e.g., asking students to revise reviews where all comments already con-

using development data from prior classes.

[5] Starting with unigrams gave us a noisy set and degraded model performance. We plan to apply LDA (Blei et al., 2003) for this task in future.

Model	Acc.	κ	F1:Sln	F1:Prb	F1:Non
BoW	0.50	0.24	0.40	0.51	0.57
SWoRD	**0.62**	**0.44**	**0.55**	**0.59**	**0.72**

Table 1: Comment-level solution prediction performance. Acc: Accuracy, F1 by class label is reported – Sln: Solution, Prb: Problem-only, Non: Non-criticism.

tained solutions) or to incorrectly display comments as red or green in the feedback interface.

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., triggered instant feedback), 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 κ 0.87).

Only intervened reviews were used to evaluate model performance because subsequent resubmissions were not predicted. In our deployment, 134 of 1428 reviews 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[6] in 3-way classification. Given that the AP data was never used for model training, the obtained performance 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.

[6]Used 1,2,3-grams as features.

True S-RATIO	≤ 1.0	> 0.7	$= 1.0$
#intervened	134	24 (18%)	16 (12%)

Table 2: True S-RATIO of intervened submission

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 (0.9%) was 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.

Acknowledgments

This research is supported by NSF/1122504, and IES/R305A120370. The larger research project (co-led by Professors Kevin Ashley, Amanda Godley, Diane Litman and Chris Schunn) is a collaboration with the Learning Research & Development Center and the School of Education. We are grateful to our colleagues in the project.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Kwangsu Cho and Christian D. Schunn. 2007. Scaffolded Writing and Rewriting in the Discipline: A Web-based Reciprocal Peer Review System. *Computers & Education*, 48(3):409–426, April.

Andrea Ernst-Gerlach and Gregory Crane. 2008. Identifying Quotations in Reference Works and Primary Materials. In *Research and Advanced Technology for Digital Libraries*, volume 5173 of *Lecture Notes in Computer Science*, pages 78–87. Springer Berlin Heidelberg.

Sarah Gielen, Elien Peeters, Filip Dochy, Patrick Onghena, and Katrien Struyven. 2010. Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, 20(4):304 – 315. Unravelling Peer Assessment.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically Assessing Review Helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristi Lundstrom and Wendy Baker. 2009. To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, 18(1):30–43, March.

Melissa Nelson and Christian Schunn. 2009. The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, 37(4):375–401.

Huy Nguyen and Diane Litman. 2013. Identifying Localization in Peer Reviews of Argument Diagrams. In *Artificial Intelligence in Education*, volume 7926 of *Lecture Notes in Computer Science*, pages 91–100. Springer Berlin Heidelberg.

Huy Nguyen and Diane Litman. 2014. Improving Peer Feedback Prediction: The Sentence Level is Right. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 99–108, Baltimore, Maryland, June. Association for Computational Linguistics.

Huy Nguyen, Wenting Xiong, and Diane Litman. 2014. Classroom Evaluation of a Scaffolding Intervention for Improving Peer Review Localization. In *Intelligent Tutoring Systems*, volume 8474 of *Lecture Notes in Computer Science*, pages 272–282. Springer International Publishing.

Lakshmi Ramachandran and Edward F. Gehringer. 2013. A User Study on the Automated Assessment of Reviews. In *Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education AIED 2013, Memphis, USA, July 9-13, 2013*.

Lakshmi Ramachandran and Edward F. Gehringer. 2015. Identifying Content Patterns in Peer Reviews Using Graph-based Cohesion. In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2015, Hollywood, Florida. May 18-20, 2015.*, pages 269–275.

Wenting Xiong and Diane Litman. 2010. Identifying Problem Localization in Peer-Review Feedback. In *Intelligent Tutoring Systems*, volume 6095 of *Lecture Notes in Computer Science*, pages 429–431. Springer Berlin Heidelberg.

Wenting Xiong, Diane J Litman, and Christian D Schunn. 2010. Assessing Reviewer's Performance Based on Mining Problem Localization in Peer-Review Data. ERIC.

Wenting Xiong, Diane Litman, and Christian Schunn. 2012. Natural Language Processing techniques for researching and improving peer feedback. *Journal of Writing Research*, 4(2):155–176. Query date: 2015-05-24.