Classroom Evaluation of a Scaffolding Intervention for Improving Peer Review Localization

Huy Nguyen, Wenting Xiong, and Diane Litman

Department of Computer Science,
University of Pittsburgh, Pittsburgh, PA 15260

Abstract. A peer review system that automatically evaluates student feedback comments was deployed in a university research methods course. The course required students to create an argument diagram to justify a hypothesis, then use this diagram to write a paper introduction. Diagram and paper first drafts were both reviewed by peers. During peer review, the system automatically analyzed the quality of student comments with respect to localization (i.e. pinpointing the source of the comment in the diagram or paper). Two localization models (one for diagram and one for paper reviews) triggered a system scaffolding intervention to improve review quality whenever the review was predicted to have a ratio of localized comments less than a threshold. Reviewers could then choose to revise their comments or ignore the scaffolding. Our analysis of data from system logs demonstrates that diagram and paper localization models have high prediction accuracy, and that a larger portion of student feedback comments are successfully localized after scaffolded revision.

Keywords: Peer review, review localization, scaffolding, evaluation.

1 Introduction

While peer review is a promising approach for helping students improve their writing, peer feedback can be of mixed quality. For example, prior work [6,5] has shown that feedback is more likely to be implemented in a revision when the review is *localized*, that is, pinpoints the location of the problem mentioned in the feedback (as shown in the examples in Fig. 1). As a first step towards helping students improve the quality of their feedback, natural language processing and machine learning have been used to build models for automatically detecting whether peer reviews contain localization and other desirable feedback properties [2,11,8,7]. To date, however, such models have typically been evaluated only intrinsically (i.e. with respect to predicting gold standard manual annotations), rather than extrinsically with respect to a real-world task (e.g. being incorporated into a peer review system to improve review quality). In addition, while intrinsic evaluations have shown that a predictive model can yield high accuracy when trained and tested on data from the same peer-review assignment, how the model performs on unseen data sets has not yet been examined. To address these

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 272–282, 2014.

issues, we have conducted both an intrinsic and extrinsic evaluation of review localization in a classroom setting. First, we followed our previous work [11,7] to implement models for predicting localization in comments of paper and diagram reviews, and integrated them into SWoRD [3], a web-enabled peer review system.[1] Next, we designed and implemented a system scaffolding intervention to improve students' use of localization when they provide feedback to each other. In our intervention, scaffolding is triggered whenever a review is predicted to have a ratio of localized comments less than a threshold. Students (as reviewers) can then choose to either revise their review comments or ignore the scaffolding. Finally, we deployed this system in a classroom setting, and evaluated its success from several perspectives. Our results show that for both diagram and paper reviews 1) the localization models predict the absence of localization in reviews with high accuracy, 2) the system scaffolding intervention helps reviewers to revise their feedback to increase localization, and 3) reviewers continue to add localization even after the scaffolding is removed.

2 Related Work

In instructional science, Gielen et al. [1] investigated effects of different peer feedback characteristics and showed that the presence of feedback justification significantly improved writing performance. Nelson and Schunn [6] found that localization in reviews of papers was significantly related to problem understanding, which in turn was significantly related to feedback implementation. Lippman et al. [5] similarly showed that localization was related to the implementation of peer feedback on argument diagrams.

Based on findings such as the above, research in computer science has used natural language processing and supervised machine learning to automatically detect when a free text feedback comment exhibits a desirable quality. Xiong and Litman [11] developed models for predicting localization in peer reviews of written papers, using features derived from a dependency parse tree. Nguyen and Litman [7] developed a localization model tailored to reviews of diagrams rather than papers, by considering common words between review comments and the target diagram.

Similar methods have been used to predict feedback helpfulness label (Yes v. No) [2], helpfulness rating [12], and other measures of review quality [8]. Particularly, we found in our prior work [12] that the percentage of localized comments contributes to improving performance of modeling review helpfulness. In this paper, instead of developing new prediction models, we focus on integrating existing models of review localization into a working peer review system, and evaluating model performance in a classroom deployment.

[1] While it is possible to modify a reviewing interface to have reviewers directly comment upon a paper, such an interface encourages primarily feedback on low-level text issues, and is not good for repeated errors or issues with larger sections of text. Therefore, we focus on encouraging localization in end-comments.

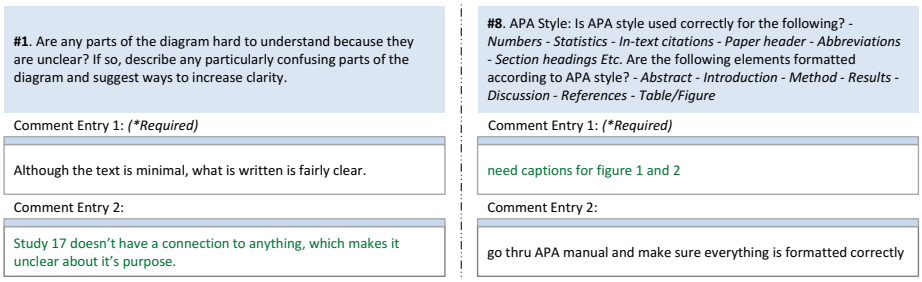

Fig. 1. Examples of localized (in green) and not localized (in black) comments in a diagram review (left) and a paper review (right). Localization cues in the green comments are *"Study 17"* (left) and *"figure 1 and 2"* (right).

Regarding system scaffolding to increase feedback quality, the design of our intervention incorporates techniques from prior work in intelligent tutoring systems. Razzaq and Heffernan [9] compared two approaches for giving hints during tutoring: proactively when students make errors, versus on-demand when students ask for a hint. They found no difference in learning gains for students who did not ask for many hints. Because our students are not trained on feedback localization we do not expect them to know when they need a hint, and thus choose to trigger our scaffolding intervention proactively whenever a student review lacks sufficient localization. In a different context, Kumar [4] showed that when error-flagging was provided during tests on introductory programming concepts, student scores improved. To implement error-flagging, correct student answers were displayed in green and incorrect answers were displayed in red; in addition, no reasons why the answers were incorrect were provided. In our system we will similarly display localized versus not-localized feedback predictions using different colors, to help students identify the problematic comments.

3 Adding Localization Scaffolding to Peer Review

A typical peer review exercise using SWoRD involves three main phases: 1) student authors create first drafts[2], 2) peer reviewers provide feedback[3] on the drafts, and 3) authors revise their drafts to address the feedback. The original version of SWoRD only facilitates the document management and review assignment aspects of peer review. To further enhance the utility of SWoRD, in this paper we add artificial intelligence to the system by integrating the detection and scaffolding of localization into phase 2, using prior models from the literature to predict paper [11] and diagram [7] review localization, respectively.

In our enhanced version of SWoRD, whenever an argument diagram or paper review is submitted, the corresponding review localization model is first used

[2] A draft can be a paper, a diagram, a presentation, etc. depending on the assignment.
[3] Feedback is in the form of written comments along with numerical ratings.

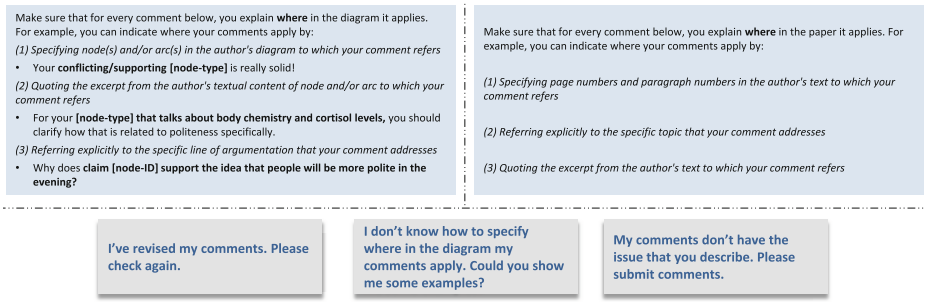

Fig. 2. Scaffolding messages for revising reviews of diagrams (top left) and papers (top right), along with the three responses available to reviewers (bottom)

to predict whether every review comment is localized or not. Fig. 1 shows examples of localized and not-localized comments from both a diagram (left) and paper (right) review, in which comments predicted as localized are highlighted in green. Then, if the submitted review is predicted to have a ratio of localized comments less than a threshold of 0.5[4], the scaffolding intervention will be triggered: the system displays an on-screen message which suggests review revision and provides advice for doing so (see the top of Fig. 2 for diagrams (left) and papers (right)). Finally, the reviewer can choose to revise the review and resubmit, view some model comments, or submit the review without revision (implying disagreement) as indicated by the three buttons at the bottom of Fig. 2. Every revised review then goes through the same localization prediction process.

4 The Peer Review Corpus

Our corpus consists of comments from *diagram* and *paper* reviews, collected from undergraduate Research Method course in psychology at University of Pittsburgh, 2013. In this class, students were asked to first create graphical argument *diagrams* using LASAD [10] to justify given hypotheses. Student argument diagrams were then distributed via SWoRD to 4 randomly assigned peers for reviewing. Student authors could revise their argument diagrams based on peer feedback, then used the diagrams to write the introduction of associated *papers*. Similarly to the diagram review step, student papers were randomly assigned to 4 peer reviewers (potentially different than the diagram reviewers). Finally, after receiving reviews of their papers, authors could revise their papers before final submission. Diagram and paper reviews both consisted of multiple feedback comments written in response to rubric prompts (e.g. #1 and #8 in the top boxes in Fig. 1). Reviewers were required to provide feedback for 5 argument diagram prompts and 8 paper prompts. Each prompt required reviewers to provide one to three comments. The system allows reviewers to edit and resubmit

[4] The threshold was tuned based on data from prior classes.

Table 1. Peer review data statistics. All re-submissions are counted.

	Diagram review	Paper review
Reviewers/Authors	181/185	167/183
Submitted reviews	788	720
Intervened submissions	173	51

Table 2. Localization annotation results

	Diagram review	Paper review
`Localized` comments	449	347
`NOT Localized` comments	718	336

their reviews at any time before the deadline with the same review scaffolding procedure. Table 1 summarizes the dataset.

To support the evaluations described below, we collected all diagram and paper review submissions which triggered a system intervention, as well as their subsequent resubmissions (if any), and then manually coded the collected reviews (both submissions and resubmissions) for the presence of localization in each comment. In addition, since reviewers may edit their submitted reviews without any system intervention, we also collected and coded localization for all reviews where re-submission occurred after a non-scaffolded submission. By pairing each comment with its revision, we aim to evaluate how the system scaffolding impacted reviewer revisions.

Following the localization annotation scheme of [5], a comment is coded as `Localized` if it contains at least one text span indicating where in the target diagram or paper the comment is applied. The comment is coded as `NOT Localized` otherwise. Two annotators independently coded comments of diagram reviews and achieved inter-rater Kappa of 0.8. The two annotators then resolved label disagreements to obtain the final labels used for our evaluations. Another annotator who had Kappa of 0.8 when coding prior paper review data was chosen to code the paper review comments obtained during our experiment. Table 2 summarizes the annotated data used in our analyses.

5 Review Localization Prediction Performance

Our first analysis aims to evaluate both the accuracy of predicting localization at the comment level, and the accuracy of using these predictions to intervene at the review submission level, for both diagram and paper reviews.

At the comment level, we evaluate how well the two review localization models predict the presence of localization compared to the manual annotations. We also compare the models' performance to their corresponding majority-class

Table 3. Localization prediction performance at the comment level

	Diagram review			Paper review		
	Accuracy	F-measure	Kappa	Accuracy	F-measure	Kappa
Baseline	61.5%	0.47	0	50.8%	0.34	0
Model	81.7%	0.82	0.62	72.8%	0.73	0.46

Table 4. Intervention prediction performance at the review submission level

	Diagram review	Paper review
Total scaffolding interventions	173	51
Incorrectly triggered scaffolding interventions	1	0

baselines[5]. Table 3 shows that both localization models substantially outperform their respective baselines. In addition, when comparing these results with the originally reported results for these models (accuracy and Kappa figures of 83.8% and .56 for diagrams [7], and 77.4% and .55 for papers [11], respectively), we see that performance is only slightly degraded in our cross-domain evaluation setting. Our current evaluation setting is more difficult because the localization models were trained prior to our corpus collection while each of the models in the original publications were trained and tested on a single dataset using cross-validation.

At the review submission level, we consider an intervention to be correct when at least one of the comments in a submission is labeled as `NOT Localized`, as reviewers should only think the system incorrectly intervened when all of the comments in a submitted review were indeed localized. As shown in Table 4, the diagram review localization model yielded only one incorrect intervention, while the system never incorrectly intervened when scaffolding a paper review.

In sum, our results show that in a real classroom setting, our models accurately predict localization in the review comments of both diagrams and papers. These comment-level predictions, in turn, are the basis of a system scaffolding intervention that is accurately triggered from a reviewer's perspective.

6 Reviewer Responses to the System Intervention

In this section, we first analyze whether reviewers actually revise their comments in response to the system scaffolding intervention. For those reviews that are indeed revised, we then analyze whether the number of localized comments in fact increases after review revision, and whether revision behavior varies depending on whether the review revision was scaffolded versus unscaffolded.

Reviewer Response Types. A reviewer can respond to the system's scaffolding intervention in one of three ways (recall the buttons shown in Fig. 2):

[5] The majority class is `NOT Localized` for diagram and `Localized` for paper review.

Table 5. Percentage of different types of reviewer responses to first interventions

Response type	`Revise`		`Disagree`		(`View Example`)	
Diagram review	54	48%	59	52%	(5)	(4%)
Paper review	13	30%	30	70%	(1)	(2%)

Table 6. Histogram of responses by true localization ratios in diagram reviews and paper reviews. NA means the bin has no data.

Ratio bin	[0,.1)	[.1,.2)	[.2,.3)	[.3,.4)	[.4,.5)	[.5,.6)	[.6,.7)	[.7,.8)	[.8,.9)	[.9,1)	1
	Diagram reviews										
Tot. responses	12	8	32	5	28	9	16	1	1	NA	1
%`Disagree`	75.0	37.5	50	20	50	77.7	43.7	0	100	NA	100
	Paper reviews										
Tot. responses	3	4	3	4	5	7	12	5	NA	NA	NA
%`Disagree`	100	50	66.7	75	60	85.7	66.7	60	NA	NA	NA

- `Revise`: the reviewer resubmits her review after revising it.
- `View Example`: the reviewer views examples of localized comments, then goes back to the system intervention interface.
- `Disagree`: the reviewer submits her review without revision.

For this paper, we consider only reviewer responses after the system's first intervention for a review.[6] Table 5 shows the percentage of different response types to these first interventions. In addition, as `View Example` is not an action that completes the review activity, the response must be followed by either a `Revise` or a `Disagree`. The number of `Revise` and `Disagree` responses thus include the responses that happened after `View Example`. As shown in Table 5, despite the system's high level of intervention accuracy (recall Table 4), reviewers disagreed more than they agreed with the system's scaffolding feedback, for both diagram and paper reviews. To investigate whether student reviewers were disagreeing with the system for good reasons (e.g., while not perfect, their review was already highly localized), Table 6 reports the percentage of the total number of responses (revisions plus disagreements) that were disagreements, with respect to different bins of true localization ratios. Pearson correlation tests between the percentage of `Disagree` responses (scaled to [0,1]) and the true localization ratio show no significant correlations (p-value's of 0.38 and 0.5 for diagram and paper review data, respectively). Student disagreement thus does not seem to be related to how well the original review had localized comments.

[6] Our data shows that first interventions account for 65% and 84% of total diagram and paper review interventions, respectively, and that reviewers were more reluctant to edit their comments in resubmissions. Based on these findings, the current version of SWoRD has been revised to intervene only once.

Table 7. Comment change patterns by intervention scopes

Change pattern	Scope=In		Scope=Out		Scope=No	
	Number of comments of diagram reviews					
NOT Localized → Localized	26	30.2%	**7**	**87.5%**	3	12.5%
Localized → Localized	26	30.2%	1	12.5%	**16**	**66.7%**
NOT Localized → NOT Localized	**33**	**38.4%**	0	0%	5	20.8%
Localized → NOT Localized	1	1.2%	0	0%	0	0%
	Number of comments of paper reviews					
NOT Localized → Localized	8	20%	**2**	**50%**	5	9.1%
Localized → Localized	13	32.5%	1	25%	**29**	**52.7%**
NOT Localized → NOT Localized	**19**	**47.5%**	1	25%	20	36.4%
Localized → NOT Localized	0	0%	0	0%	1	1.8%

Review Revision. Next we evaluate the effectiveness of the system scaffolding intervention by looking at the human-coded localization annotations for edited comments of different types, where the types are defined in terms of the prior system scaffolding interventions that a reviewer received. A reviewing session starts when the reviewer creates/opens a review and ends when the reviewer submits the review by either passing the localization threshold or disagreeing with the system (by clicking on the rightmost button in Fig. 2). We define three intervention scopes with respect to reviewer edits during a reviewing session:

- Scope=In: the reviewer received a system intervention in the current reviewing session.
- Scope=Out: the reviewer did not receive a system intervention when submitting a review for the current diagram/paper, but encountered a system intervention for a prior review of that type.
- Scope=No: the reviewer of a diagram/paper never received a system intervention for either the current or prior reviews of a diagram/paper.

For each intervention scope, we collect all comments that were edited in the revision and compare each comment's true localization label to the true label of its previous version. Table 7 reports the number of comment pairs according to the four possible ways in which a comment could be changed after editing, with respect to localization. The pattern of most interest is NOT Localized → Localized, as this was the type of successful edit that the scaffolding intervention was designed to promote. At the other extreme, the least desirable pattern is Localized → NOT Localized, as this type of comment editing decreased feedback quality with respect to localization.

First, consider the first rows for both the diagram and paper reviews in Table 7, which correspond to the most desirable edit pattern. Comparing columns shows that the percentages of NOT Localized → Localized in Scope=In and Scope=Out are larger than that of Scope=No, for both diagram and paper re-

views. Moreover, in `Scope=Out` this pattern contributes the largest portion of edits in both diagram and paper review revisions. Such evidence indirectly suggests that the system scaffolding intervention does help reviewers to localize their previously unlocalized comments, and the impact of the intervention still remains in later reviewing sessions after the scaffolding is removed.

The second pattern in the table, `Localized` → `Localized`, has the largest percentage in `Scope=No`. We hypothesize that reviewers who were never scaffolded might be revising their reviews for some reason other than feedback localization which they already did well. However, this pattern also contributes the second largest percentages for the other two scopes. Perhaps reviewers might also be attempting to add more localization signals than that were used in their original comments. In future work we plan to revisit our localization coding (which currently has a binary rather than ordinal value) to determine whether reviewer editing adds further localization, or addresses a different issue.

Our third observation is that for `Scope=In`, the pattern `NOT Localized` → `NOT Localized` accounts for the largest number of edit results in both diagram and paper reviews. This suggests that there is still room for improvement in our scaffolding of review localization. That is, even when reviewers attempted to respond to the system intervention by revising their comments and asking the system to evaluate them again, students still had difficulty in making the comments localized. Potential reasons might be that our current scaffolding messages could be made clearer, or that for some review dimensions giving localized comments is difficult. Investigating these issues will be part of our future work.

Finally, the least desirable pattern of `Localized` → `NOT Localized` occurred only twice in all of the edits. We investigated these instances and found that students apparently deleted their comments by mistake. The rareness of this pattern suggests that our highlighting of localized comments in green helped student reviewers not to remove localization from their localized comments.

7 Conclusions and Future Work

In this paper, we first integrated two review localization models for diagram and paper reviews in a web-based peer review system, then implemented a scaffolding intervention to improve the quality of peer reviews that lacked localization. Furthermore, we deployed the system in a university classroom and evaluated the system in terms of the prediction performance of the two localization models (in a cross-domain fashion), the system scaffolding intervention triggered by these models, and the effect of scaffolding on reviewer revision behavior, using data from the class. Our comment-level results showed that both localization models outperformed majority class baselines, with absolute performance levels approaching prior laboratory results [11,7]. Our review submission-level results demonstrated that the two localization models could also accurately trigger system interventions, yielding only one wrong intervention for a diagram review.

Analyzing reviewer responses to the system intervention, we found that for reviewers who revised their reviews after the system scaffolding intervention,

the number of comments with localization increased after editing. Moreover, the scaffolding intervention appeared to improve localization even in later, non-scaffolded review sessions. However, the results also demonstrated that our current approach could be further improved, as there were both a large number of unsuccessful attempts to localize comments, and a large number of disagreements with the system's suggestion to increase localization.

For future work, we plan to improve our interface to better help students localize their review comments. In addition to using color to distinguish localized and non-localized comments, we plan to highlight the localized text spans in already localized comments (e.g. "Study 17" in the left of Fig. 1). We also plan to do further annotation to examine not only whether, but how strongly, a comment is localized. Finally, we plan to ask reviewers why they are disagreeing with the system, as our initial analyses did not show any relationship with localization.

Acknowledgments. This research is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120370 to the University of Pittsburgh. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. Our work is also supported by NFS 1122504. We are grateful to our colleagues for sharing the data. We thank M. Lipschultz, K. Ashley, C. Schunn and other members of the ArgumentPeer and ITSPOKE groups as well as the anonymous reviewers for their valuable feedback.

References

1. Gielen, S., Peeters, E., Dochy, F., Onghena, P., Struyven, K.: Improving the effectiveness of peer feedback for learning. Learning and Instruction 20(4), 304–315 (2010)
2. Cho, K.: Machine Classification of Peer Comments in Physics. In: Proceedings of 1st international conference on Educational Data Mining (EDM), pp. 192–196 (2008)
3. Cho, K., Schunn, C.D.: Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. Computers & Education 48(3), 409–426 (2007)
4. Kumar, A.N.: Error-Flagging support for testing and its effect on adaptation. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 359–368. Springer, Heidelberg (2010)
5. Lippman, J., Elfenbein, M., Diabes, M., Luchau, C., Lynch, C., Ashley, K.D., Schunn, C.D.: To Revise or Not To Revise: What Influences Undergrad Authors to Implement Peer Critiques of Their Argument Diagrams? In: International Society for the Psychology of Science and Technology 2012 Conference, Poster (2012)
6. Nelson, M.M., Schunn, C.D.: The nature of feedback: How different types of peer feedback affect writing performance. Instructional Science 37(4), 375–401 (2009)
7. Nguyen, H.V., Litman, D.J.: Identifying Localization in Peer Reviews of Argument Diagrams. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS, vol. 7926, pp. 91–100. Springer, Heidelberg (2013)
8. Ramachandran, L., Gehringer, E.F.: Automated assessment of review quality using latent semantic analysis. In: 11th IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 136–138 (2011)

9. Razzaq, L., Heffernan, N.T.: Hints: is it better to give or wait to be asked? In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 349–358. Springer, Heidelberg (2010)
10. Scheuer, O., Loll, F., Pinkwart, N., McLaren, B.M.: Computer-supported argumentation: A review of the state of the art. International Journal of Computer-Supported Collaborative Learning 5(1), 43–102 (2010)
11. Xiong, W., Litman, D.: Identifying problem localization in peer-review feedback. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part II. LNCS, vol. 6095, pp. 429–431. Springer, Heidelberg (2010)
12. Xiong, W., Litman, D.: Automatically Predicting Peer-Review Helpfulness. In: Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), pp. 502–507 (2011)