Int J Artif Intell Educ (2017) 27:582–622
DOI 10.1007/s40593-016-0136-6

ARTICLE

Iterative Design and Classroom Evaluation of Automated Formative Feedback for Improving Peer Feedback Localization

Huy Nguyen[1] · **Wenting Xiong**[2] · **Diane Litman**[3]

Published online: 6 January 2017

Abstract A peer-review system that automatically evaluates and provides formative feedback on free-text feedback comments of students was iteratively designed and evaluated in college and high-school classrooms. Classroom assignments required students to write paper drafts and submit them to a peer-review system. When student peers later submitted feedback comments on the papers to the system, Natural Language Processing was used to automatically evaluate peer feedback quality with respect to localization (i.e., pinpointing the source of the comment in the paper being reviewed). These evaluations in turn triggered immediate formative feedback by the system, which was designed to increase peer feedback localization whenever a feedback submission was predicted to have a ratio of localized comments less than a threshold. System feedback was dynamically generated based on the results of localization prediction. Reviewers could choose to either revise their feedback comments to address the system's feedback or could ignore the feedback. Our analysis of data from system logs demonstrates that our peer feedback localization prediction

✉ Huy Nguyen
hvn3@pitt.edu

Wenting Xiong
wenting.pitt@gmail.com

Diane Litman
dlitman@pitt.edu

[1] Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260, USA

[2] IBM Watson, Yorktown Heights, NY 10598, USA

[3] Department of Computer Science and Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA 15260, USA

model triggered the formative feedback with high precision, particularly when peer feedback comments were written by college students. Our findings also show that although students often incorrectly disagree with the system's feedback, when they do revise their peer feedback comments, the system feedback was successful in increasing peer feedback localization (although the sample size was low). Finally, while most peer comments were revised immediately after the system feedback, the desired revision behavior also occurred further after such system feedback.

Keywords Peer feedback · Feedback localization · Automated formative feedback

Introduction

A typical peer assessment practice when learning to write is asking students to reciprocally review other students' work and generate peer feedback. Peer feedback is an important alternative to teacher feedback and is used frequently because it enhances students' learning by giving them learning opportunities in their roles as both author and reviewer while not increasing teacher workload (Kern et al. 2003; Cho et al. 2007; Lundstrom and Baker 2009; Cho and MacArthur 2011; Nicol et al. 2014). In the domain of writing, peer feedback is usually referred to as "peer review" or "peer assistance when writing" (Gielen et al. 2010). Peer feedback can take many forms: it may be face-to-face or written feedback, and feedback may involve numeric ratings, free-text comments or both. Our current research focuses on peer feedback with free-text comments.

While peer feedback with free-text comments is a promising approach for helping students improve their writing, feedback comments from peer reviewers can be of mixed quality (Nilson 2003; Nelson and Schunn 2009; Cho and MacArthur 2010; Gielen et al. 2010). Prior work has operationalized feedback quality from different perspectives: (1) accuracy and consistency across reviewers and/or with teacher's feedback (Steendam et al. 2010), (2) content and/or style characteristics (see the work by Gielen et al. (2010) for a review of prior research from this perspective). The advantage of the second approach to feedback quality is that the proposed feedback characteristics are task-independent, thus the acquired peer assessment skills are transferable to other settings. While feedback characteristics were usually derived from expert assessment reports and grounded in learning theories (Gielen et al. 2010), follow-up research was conducted to empirically investigate their contribution to writing performance. For example, Nelson and Schunn (2009) and Lippman et al. (2012) showed that feedback comments are more likely to be implemented in a paper revision when the comments are localized, that is, pinpoint the location of the problem mentioned in the feedback. Gielen et al. (2010) found that feedback comments with justification, that is, include an explanation of judgment, significantly improve writing performance.

As the first steps towards helping student reviewers improve the quality of their peer feedback, Natural Language Processing and Machine Learning have been used to build models for automated peer feedback assessment (Cho 2008; Xiong and Litman 2010; Ramachandran and Gehringer 2011; Nguyen and Litman 2013). For

example, Ramachandran and Gehringer (2013) developed a model to automatically provide formative assessment of peer feedback on different metrics such as review content type and review tone type. In similar veins of study, research has been conducted to automatically detect whether feedback comments lack localization or other desirable properties (Xiong et al. 2010; Nguyen and Litman 2013; 2014). To date, however, while models for assessing such properties of peer feedback have been evaluated intrinsically (i.e., with respect to predicting gold standard manual annotations), extrinsic evaluation of their application in real-world tasks (e.g., being incorporated into a peer-review system to improve peer feedback quality) has been studied limitedly. To the best of our knowledge, the work by Ramachandran and Gehringer (2013) was the only research on evaluating an automated peer feedback assessment system. However, their system only provided feedback to student reviewers at the end of the peer feedback process, and no analysis of the impact of the system's evaluative feedback on peer feedback quality was conducted. Currently no research has investigated peer-review systems that provide interactive feedback to help reviewers improve their comments within the peer feedback process. On the other hand, while intrinsic evaluations have shown that models for automatically evaluating peer feedback quality can yield high accuracy when trained and tested on peer comments from the same writing assignment, the performance of such models on more challenging and realistic types of test data (e.g., from different writing assignments, academic disciplines, student grade levels) has typically not been examined.

To address these issues, we enhanced a peer-review system by developing an *automated formative feedback* strategy that uses Natural Language Processing to automatically evaluate the quality of peer feedback with respect to localization, then uses these evaluations to trigger dynamically-generated formative feedback designed to improve peer feedback localization. Our peer feedback localization prediction model processes all peer feedback comments as they are submitted to the peer-review system, and provides real-time feedback to peer reviewers indicating whether their comments were identified as localized or not. Student reviewers can choose to revise their comments regarding feedback localization and resubmit their comments, or ignore the system's feedback and submit the original peer comments. We intrinsically and extrinsically evaluated the automated formative feedback strategy on four writing assignments, using peer comment data from system deployments in one high-school Math and two college Psychology classes. The goal of our comprehensive evaluation of a peer-review system is to find answers to the following three research questions:

1. *How precisely does our automated formative feedback strategy evaluate peer feedback localization?* An insight of prediction accuracy will help us further improve system performance.
2. *Are student reviewers likely to agree or disagree with system's formative localization feedback?* This question helps us understand reviewer behaviors toward the system feedback, and how such responses relate to the system performance.
3. *How does the system's feedback impact peer feedback revision?* In other words, we want to know whether the designed formative feedback helps student reviewers improve their feedback comments with respect to localization. Answering

this question will give us insight on student learning and its relation with interface design to further optimize the system.

Feedback Localization

In the domain of writing, feedback localization is defined as pinpointing the source or location with respect to the paper being reviewed of the issue being discussed in the feedback comments (Nelson and Schunn 2009). For an operational definition of feedback localization that supports our prediction model, we consider three types of localization. For each type of localization, example comments are provided in italic text, and localization text is in boldface.

1. Explicit localization by position: specifying the position in the reviewed paper, using absolute page, paragraph, and/or sentence numbers, a relative positional expression, or section heading:
   - *Your entire hould take up **page 2**, therefore, the **Intro** should start on **page 3** and should not say introduction, rather should say your title (APA)*
   - *In the **fifth paragraph** on the **first page**, you say "holds much importance". This just sounds awkward and should be revised.*
2. Implicit localization by content/topic: referring to the content/topic of the reviewed aspect of the paper:
   - *I would check the sentence where you begin to talk about **perceiving and discriminating things based on features a look**.*
   - *In the third paragraph you give the reader an idea of what **featural processing and configural processing** is, it would be helpful to do this earlier in the paper.*
3. Quoted text: quoting excerpts from the reviewed paper:
   - *You had a few instances of awkard wording throughout your paper. For example, in the third paragraph, you say, **"The mothers' ages were around 37."***
   - *You have thrown me off a bit when you said **"The rats will stay alive the whole first year they are there"**. Why and where are they? This is a detail to include when nothing or nobody will go missing.*

It is possible that the reviewer refers to the text of the entire document, e.g., "*The biggest problem was grammar and punctuation*", but we find this type of document-level location information has very limited benefit to the student authors. Our student authors, especially high-school or lower-grade students, are just novice writers with very little writing experience. Peer feedback that mentions a general issue without any particular example makes them confused in understanding the issues and revising their writing. In fact, our annotated data shows evidence that comments with only high-level localization received a very low implementation rate. Therefore, we do not consider a reference to the entire text of the document as a localization in this study.

Paper Structure

The rest of this paper is organized as follows. The next section discusses work that motivated our current research, ranging from educational to technical aspects. We then describe the SWoRD peer-review system as a software platform for developing and evaluating our automated formative feedback strategy for improving peer feedback localization. Next, we split our research into three studies and describe them in three following sections, respectively. Our first study (Study 1 section) focuses on the implementation of the automated formative feedback strategy and addresses our first research question. We describe the training data and Machine Learning algorithm used to build the peer feedback localization prediction model. We then present prediction performance at both peer feedback comment and peer feedback submission levels for different classes. Our second study (Study 2 section) analyzes data collected from different deployments of the automated formative feedback strategy to evaluate student reviewer responses to the system's feedback, i.e., agree vs. disagree, and addresses our second research question. Our third study (Study 3 section) evaluates the impact of the system's feedback on feedback revision by student reviewers, which addresses the third research question. The last two studies thus investigate reactions of student reviewers to the system's formative feedback in terms of whether they agreed or disagreed with the system feedback, and how they revised their comments when they agreed with the system, respectively. The final section of this paper presents a concluding discussion of the contributions, limitations and implications of our research.

Related Work

Improving Peer Feedback Quality

In educational contexts, peer assessment has been shown to help learners improve self-evaluation skills and better understand concepts being studied. As a result, the practice is being used with increasing frequency across disciplines, especially in content-area courses (Kern et al. 2003; Topping 2009; Lundstrom and Baker 2009). Different aspects of peer assessment such as learners' perceptions (Mulder et al. 2014), impact on revision (Kaufman and Schunn 2011), design principles and effective uses (Berg et al. 2006; Landry et al. 2014), and cognitive processes of reviewing activities (Nicol et al. 2014) have been studied to best promote student learning through the practice.

Regarding peer feedback in the form of free-text comments, studies have empirically demonstrated important characteristics of written peer feedback that relate to writing performance and feedback implementation. Gielen et al. (2010) found that the presence of justification (i.e., explanation of judgment) significantly improved writing performance. Similarly, a study by Strijbos et al. (2010) revealed that elaborated specific feedback, e.g., feedback that addresses knowledge about concepts and mistakes, leads to improved performance and outcomes. In a different study, Nelson and Schunn (2009) argued that feedback features (e.g., summarization, specificity,

explanation, scope, and effective language) may not directly affect feedback implementation but instead impact implementation through internal mediators (i.e., the feedback receiver's understanding and agreement) because of the complex nature of writing performance. Their study found that two components of the specificity feature, which were offering a solution and localization, significantly correlated with an understanding of the problem. Understanding, in turn, was found to have a significant positive relationship with feedback implementation.

To further promote the desired quality characteristic of peer feedback, research has considered using instructions and question prompts to elicit high-quality feedback (Gan and Hattie 2014; Gielen and De Wever 2015). In a prior work, Nilson (2003) proposed a set of feedback prompts that does not ask for judgment or opinion which may evoke emotion, but requires student reviewers to attend to the details of the peer's work, which encourages feedback localization. Given both the theoretical and empirical recognition that localization is an important characteristic of written feedback, we are motivated to develop an automated formative feedback strategy aimed at increasing the amount of localization in the written comments that students produce during peer feedback.

Besides providing peer feedback in the form of end-comments as we have discussed so far, graphical feedback interfaces that allow reviewers to directly annotate papers have also been used (see the NB project - nb.mit.edu - for such a system). While such feedback interfaces inherently bind comments to problem sources in the papers, research has shown that on-paper annotations encourage primarily feedback on low-level text issues (e.g., grammar, punctuation, spelling) and lead to simple erasures rather than substantive revisions (McCarthey et al. 2013; 2014). Similar findings were found by Ellis (2011) who conducted two parallel peer feedback classes (i.e., same teacher and identical feedback instructions): one let students write feedback on the printed papers and the other used an online-blog system to edit feedback on papers. The on-paper class had a much higher incidence of surface-revision comments while the blog-based class yielded more macro structure-revision comments, e.g., including main meaning or an overall summary of the entire text. In addition, end-comment feedback has been found to be more useful when providing feedback that refers to multiple locations (Ferris et al. 2013). Although the primary goal of our current research is peer feedback localization, we also care about the overall helpfulness of peer feedback in a balance between surface-level and meaning-level comments. Therefore, we currently focus on encouraging localization in end comments to make the system simple yet robust. In the future, a peer-review system which offers a combined editing mode is worth exploring.

Automated Peer Feedback Assessment

Based on findings such as discussed above, research in Computer Science has used Natural Language Processing and supervised Machine Learning to automatically evaluate whether a free-text feedback comment exhibits a desirable quality, with downstream goals such as automatically prompting student reviewers to improve their feedback quality. Xiong and Litman (2010) developed models for predicting

localization in peer feedback comments on student papers, using features derived from regular expressions and dependency parse trees. Nguyen and Litman (2013) developed a feedback localization prediction model tailored to feedback on diagrams rather than papers, by considering common words between feedback comments and the target diagram. Similar methods have been used to identify feedback helpfulness labels (e.g., helpful versus not helpful) (Cho 2008), numeric helpfulness ratings (Xiong and Litman 2011), and other measures of feedback quality such as the presence of solutions to problems (Xiong et al. 2012; Nguyen and Litman 2014). Feedback quality has also been determined by content type (e.g., the presence of problem identification and solution suggestion), relevance, coverage, tone (e.g., positive, negative or neutral), and plagiarism (Ramachandran and Gehringer 2011; Ramachandran et al. 2016). Often these measures are not independent, e.g., we found in our prior work that the percentage of localized comments contributed to improving performance when predicting numeric helpfulness ratings (Xiong and Litman 2011). Of similar motivation but in a different perspective, Ramachandran and Gehringer (2015) created a model to identify the content type of peer feedbcak, i.e., summative, advisory, or problem identification. Their study has a potential application in incorporating peer feedback based on helpfulness and content types into automated essay assessment. In this paper, instead of focusing on developing new methods for automatically predicting the presence of peer feedback features, i.e., localization, we focus on integrating automated prediction research into a working peer-review system with the goal of improving peer feedback localization. Unlike prior work, we not only conduct traditional prediction performance evaluations with respect to predicting localization, but also evaluations focused on triggering system feedback and quantifying students' responses. To do so, we intrinsically (Study 1) and extrinsically (Studies 2 and 3) evaluate the peer-review system using peer feedback test data from multiple classroom deployments.

Use of Automated Peer Feedback Assessment

To the best of our knowledge there are only a very few studies that analyze the helpfulness of automated peer feedback assessment in working peer-review systems. Ramachandran and Gehringer (2013) conducted a small user study (24 participants) of a peer-review system that incorporated an automated peer feedback assessment feature. Their results showed that student reviewers found system feedback regarding peer feedback's content type and plagiarism to be informative. However, feedback by their system was provided to reviewers at the end of their reviewing practice. We believe interactive feedback should give student reviewers different learning opportunities. Therefore, our strategy for using automated peer feedback assessment includes analyzing every single comment of peer feedback, highlighting comments that need revision, and allowing reviewers to revise and resubmit their feedback. We perform novel evaluations examining how student reviewers respond to a system's formative feedback interactively with respect to both accepting the system's feedback (Study 2) and increasing localization in revised peer feedback (Study 3).

Formative Feedback Design

To assure the main goal of formative feedback which is to improve learning, sets of guidelines on feedback characteristics, feedback timing, and feedback delivery have been proposed (Hattie and Timperley 2007; Shute 2008; Narciss 2013). Narciss (2013) further emphasized the tutoring function of formative feedback, which tutors "students to detect errors, overcome obstacles and apply more efficient strategies for solving learning tasks." However, designing formative feedback strategies for digital learning environments is challenging, and many digital educational systems do not provide tutorial feedback strategies but simple feedback strategies offering knowledge of results and/or correct responses (Narciss 2013). In the context of our formative feedback, providing system feedback on the correctness of a learner response, i.e., presence of localization in a peer feedback comment, is challenging because the system can only rely on bounded performance of feedback localization prediction. Despite such performance limitations, our design of formative feedback used our predictions in conjunction with guidelines regarding feedback display, e.g., not-localized comment flagging, highlighting of localized comments and localization text within those comments, examples of pre-selected localized comments.

Feedback highlights and error flags have been shown to be effective in automatically tutoring students. Heift (2004) found that students are more likely to revise their mistakes when given meta-linguistic feedback highlighting the mistake than when given feedback with no highlighting but with explanation. In another study that examined feedback highlighting but in a different context, Kumar (2010) 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.

With respect to scheduling system feedback, 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 system feedback proactively whenever a student feedback lacks sufficient localization. In a prior study, Razzaq and Heffernan (2010) 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. However, it is shown that students pay more attention to immediate than delayed feedback (Van der Kleij et al. 2012).

Automated Formative Feedback Strategy in SWoRD

SWoRD Peer-review System

As a software platform for developing and evaluating our strategy of automated formative feedback for improving peer feedback localization, we use SWoRD (Scaffolded Writing and Rewriting in the Disciplines), a web-based reciprocal peer-review system (Cho and Schunn 2007). A typical peer-review writing exercise using SWoRD involves four main phases: (1) students (as authors) submit first draft papers, (2)

students (as reviewers) submit feedback in the form of end-comment reviews and numeric ratings on peers' papers, (3) students (as authors) receive peer feedback on their papers, and (4) students (as authors) submit a paper revision that addresses feedback from their peers. See Appendix B for a more detailed description with examples of SWoRD's work-flow.

The original version of SWoRD only facilitated the document management and paper-reviewer assignment aspects of the peer feedback process as described above. To further enhance the utility of SWoRD, we have implemented automated prediction of and formative feedback on peer feedback localization in phase 2 of SWoRD (i.e., peer feedback submission). The goal of our work is to improve the localization of written feedback comments on a paper before they are shown to the author, by automatically providing formative feedback to student reviewers whenever localization is poor at the time of feedback submission.

In particular, we first integrated an improved version of the feedback localization prediction model developed by Xiong and Litman (2010) into SWoRD. Next, we iteratively designed and implemented automated formative feedback to improve students' use of localization in peer feedback. The work-flow of the automated formative feedback strategy in phase 2 of SWoRD is as follows. (2.1) Whenever peer feedback is submitted, the feedback localization prediction model is first used to predict whether every feedback comment is localized or not. (2.2) If the submitted feedback is predicted to have a ratio of localized comments less than a threshold of 0.5 (justified in detail in Study 1), formative feedback will be triggered automatically. In the formative feedback, e.g., Fig. 1, the system displays an on-screen message which suggests comment revision and provides advice for doing so (part A of the figure). Comments are displayed in the text boxes (parts B, C) for the reviewer to edit. (2.3) Finally, the reviewer can choose to revise his/her feedback comments and resubmit (click the left button in part A) or to submit the peer feedback without revision (click the right button), which implies agreement or disagreement with the system's feedback, respectively.

Iterative Design of Automated Formative Feedback

We have iteratively developed the automated formative feedback strategy across two different system deployments in actual classrooms. Given the lessons learned from our first deployment in 2013 (Fig. 1), we modified the formative feedback strategy before our second deployment in 2014 (Fig. 2). While the main work-flow was the same across deployments, we changed how the system processed peer feedback submissions versus resubmissions, and also enhanced the formative feedback interface.

For both of our formative feedback strategy deployments, we define a reviewing session as a work-session that starts when the interface (see Fig. 5 in Appendix B) is opened for a reviewer to enter his/her comments, and terminates when the feedback is submitted successfully to SWoRD or the reviewer closes the interface. In the 2013 deployment, the formative feedback strategy evaluates peer feedback comments for localization both during the first feedback submission as well as during any resubmissions in a reviewing session. In that deployment, feedback comments

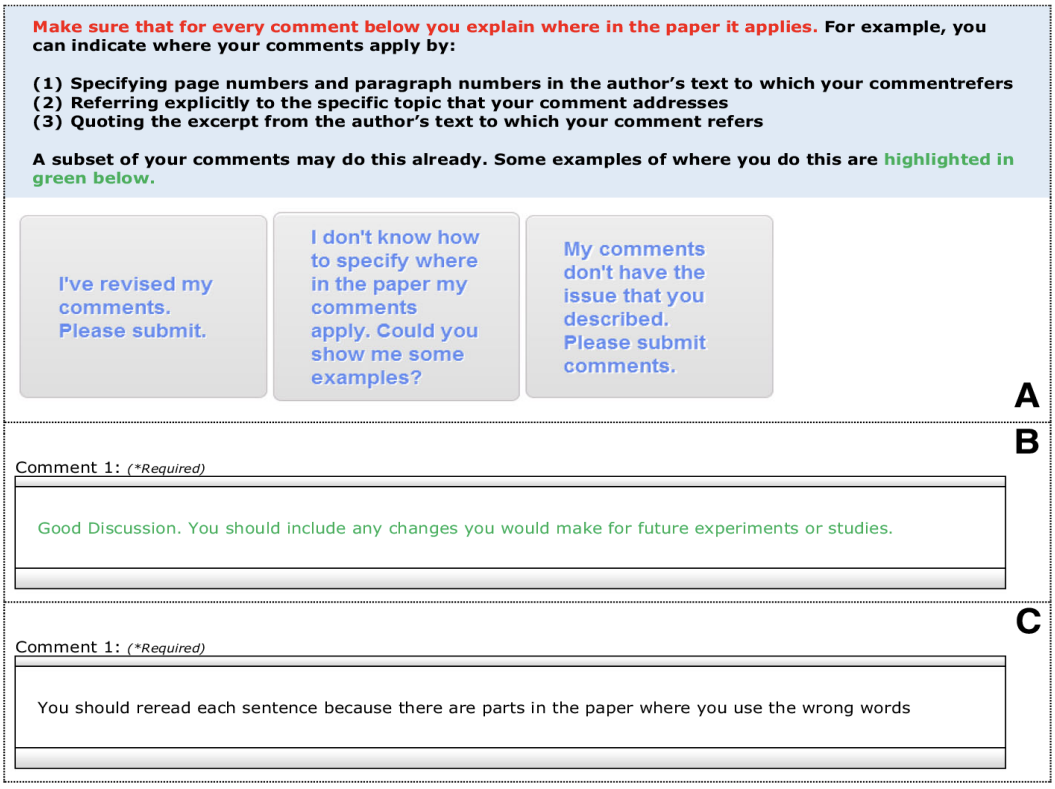

Fig. 1 Formative feedback interface added to SWoRD during the 2013 deployment of the automated formative feedback strategy. Feedback comments are from a Research Methods class in Psychology. *Part A:* the system's feedback message, and three possible student reactions which are revise and resubmit – REVISE button (*left*), view examples of localized comments – EXAMPLE button (middle), and submit revision – DISAGREE button (*right*). *Part B:* comments predicted as localized are highlighted in green. *Part C:* comments predicted as not-localized are displayed without highlighting

were submitted successfully only when the (re)submission passed the peer feedback localization check or when the reviewer clicked the DISAGREE button in the formative feedback interface. That means a student reviewer could receive more than one system feedback in a session if he/she kept failing to provide location information in comments. However, a data analysis of the 2013 deployment showed that reviewers were much more reluctant to edit their comments when the system's feedback occurred in later resubmissions. In particular, among all edits made in revisions due to system feedback, on average 12 % were from resubmissions while 88 % were made to first submissions. Also, only 17 % of comments of resubmissions that received system feedback were edited, while this ratio in first submissions was 31 %.

Therefore, in our 2014 deployment the automated formative feedback strategy only checked the first feedback submission in each reviewing session for peer feedback localization. This means that when a reviewer clicked the REVISE button, the resubmitted comments no longer went through the feedback localization prediction procedure but was instead submitted right away. Thus, while the former deployment

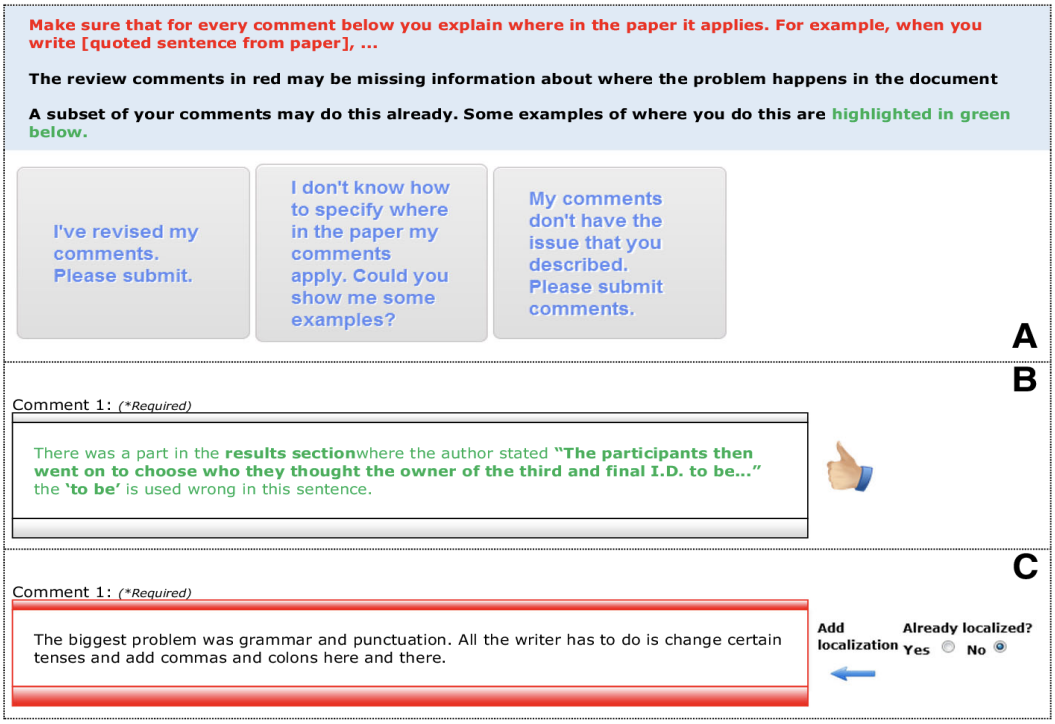

Fig. 2 Formative feedback interface added to SWoRD during the second deployment of the automated formative feedback strategy. Feedback comments are from a Cognitive Psychology class. Differences from the first deployment are as follows. *Part A:* the system's feedback message was changed to reflect the changes in Parts B and C below. *Part B:* Localized comments are also displayed with a thumbs-up icon, and the detected location signals are also highlighted in bold. *Part C:* for comments predicted as not-localized, the comment box turns red and buttons requiring reviewer feedback are now displayed

allowed many resubmissions during a reviewing session, the latter deployment allowed at most one resubmission in each session.

Along with the above difference in processing resubmissions, the 2014 deployment made changes to the formative feedback interface to better call student reviewers' attention to peer feedback localization. Unlike the original formative feedback interface (Fig. 1), the improved interface used boldface to highlight examples of localization text within the predicted localized comments (Fig. 2). This change was designed to provide student reviewers with real examples of feedback localization from their own comments. As our system now highlights localization text, we simplify the formative feedback message ("*Make sure that...*" in part A) so that it only shows a localization template (randomly chosen from three pre-defined templates - the first lines of the 3 examples in Fig. 9). In addition, the formative feedback interface includes buttons asking reviewers to provide feedback on system's feedback localization prediction (part C). This change is made to encourage students' reflection on localization. We hypothesized that asking students to reason about feedback localization in their own comments would also promote peer feedback revision. The reviewer responses to these buttons also provide new annotated examples for

supervised Machine Learning that may potentially improve the performance of our prediction model for future deployments.

Study 1: Evaluating Peer Feedback Localization Prediction

Introduction

Our first study involves two intrinsic evaluations that aim to evaluate the precision of predicting non-localization at the comment-level, and of using these predictions to trigger system feedback of the peer feedback localization. For peer feedback localization prediction, we adapt the model that was built in our initial work (Xiong and Litman 2010). Our initial set of predictive features were developed using Natural Language Processing and evaluated in a laboratory setting using cross-validation on peer feedback comments collected in a college History class. To improve model robustness for the current study, where the model was tested on data from courses for which we had no training data, we developed additional features as well as annotated more diverse training data and retrained the prediction model. In this section we describe how we extract features from peer feedback comments and associated papers. We also discuss how we bias the Machine Learning process to yield a model with high precision for triggering system feedback.

Method

Peer Feedback Data for Evaluation

In the present study, we evaluate our pre-trained feedback localization prediction models using data from two system deployments with four peer feedback assignments spanning three different classes. The first version of our feedback localization model was deployed in a Research Methods Lab at the University of Pittsburgh during Fall 2013. The second version was deployed in a Cognitive Psychology class at the University of Pittsburgh during Spring 2014, as well as in an Interactive Mathematics class at a Pittsburgh charter high school during Fall 2014. For both Research Methods and Cognitive Psychology classes, collected data consists of peer feedback comments for the first drafts of one writing assignment per class. For Interactive Mathematics class, our data includes feedback comments for the first drafts of two different writing assignments. A general description of the four peer feedback datasets is given in Table 1. Appendix A contains a detailed description for the four writing assignments in which we collect the feedback comments.

To support the evaluation of our feedback localization models and prepare data for our second and third studies, we collected all first peer feedback submissions which triggered system's feedback, as well as their immediately subsequent resubmissions (if any). By pairing each comment with its revision, we aim to evaluate how the system feedback impacted student reviewers' revision of their peer feedback. In addition, since reviewers can edit and resubmit their previously submitted feedback, we observed that there were a number of edited comments that did not immediately

Table 1 General description of peer feedback datasets

Deployment	Class	Student level	Reviewers per paper	Comment aspects
1st version	Research Methods Lab	Undergraduate	4	8
2nd version	Cognitive Psychology	Undergraduate	4	4
2nd version	Interactive Mathematics, assignment 1	High-school	3	2
2nd version	Interactive Mathematics, assignment 2	High-school	3	1

Reviewers per paper column shows the number of peer reviewers assigned to a paper, as pre-specified by the instructors. Because not all students completed all peer feedback assignments, the actual numbers of reviewers per paper range from 1 to the assigned number in every class

follow a system feedback. Thus, we also collected original and revised comments where the revision occurred without immediate system feedback, but where the reviewer had previously received the system feedback. This data enables us to explore whether there were retention effects of the system feedback for improving feedback localization.

Because peer feedback comments were collected from system's log data, each comment was associated with a predicted label of localization Thus, by comparing human-annotated labels with predicted labels, we could obtain prediction performance of our deployed prediction models. Following the localization annotation scheme used by Lippman et al. (2012), an annotator who had inter-rater Kappa of 0.8 when coding prior peer feedback data was chosen to code the collected peer feedback comments. A comment is coded as LOCALIZED if it contains at least one text span indicating where in the target paper the comment is applied. The comment is coded as NOT-LOCALIZED otherwise.

Descriptive statistics of the peer feedback data for our current study are given in Table 2. Because not all students completed every assignment and we only analyze complete reviews that were properly submitted, there are differences in the number of student authors and reviewers across assignments. As shown in the %SystemFeedback row of the table, the college classes have smaller ratios of feedback submissions that triggered system's feedback than the high-school class. In particular, both college classes have very low system feedback ratios, 7 % for Research Methods and 3 % for Cognitive Psychology. One reason could be that peer feedback prompts (see Appendix A) in the two college classes were more specific, with some targeting low-level writing issues, e.g., spelling/grammar, and others targeting particular sections in the writing, e.g., abstract, introduction. As a consequence, location information might have been more commonly added to the comments to properly address the prompts.

Prediction Features for Feedback Localization

We use 11 features to learn our feedback localization prediction model that works with different types of feedback localization that are studied in this paper (e.g., explicit localization by position in, implicit localization by content/topic of, and

Table 2 Peer feedback data

	RM	CogPsy	Math1	Math2
Reviewers	167	150	99	137
Authors	183	168	109	135
Peer Feedback Submissions	645	576	270	422
System Feedback Triggered	43	15	58	74
%SystemFeedback	7 %	3 %	21 %	18 %
Collected comments	385	133	150	77
LOCALIZED comments	185	15	71	18
NOT-LOCALIZED comments	200	118	79	59
%NOT-LOCALIZED	52 %	89 %	53 %	77 %

Peer Feedback Submission only counts first submissions. Collected comments include those in submissions that either triggered system feedback or were a revision of a previous peer feedback submission; the two conditions are not mutually exclusive. Table columns: RM – Research Methods class, CogPsy – Cognitive Psychology class; Math1 and Math2 – Interactive Mathematics class, assignment 1 and 2, respectively

quoted text from the reviewed paper – see our definitions in Feedback Localization Section). Out of the 11 features, 6 were from our laboratory study of feedback localization prediction (Xiong and Litman 2010). In particular, the regular expression feature was designed to model explicit localization by position, while the domain word and overlapping window features were designed to model the content/topic and quoted text types of localization. We, however, eliminated the fourth group of syntactic features proposed in the prior work, as performing syntactic parsing led to an unacceptable computation time in our deployment.

- **Word count**: number of words in the comment.
- **Quoted word counts**: number of words in quoted text in the comment. Quotations in comments are recognized by the occurrences of double-quote symbol (e.g., *you say, "The mothers' ages were around 37."*).
- **Comment order**: order of the comment in the feedback.
- **Regular expression tag**: a Boolean feature that indicates whether any of a predefined set of regular expressions (e.g., *on page 5*, *the section about*) are matched in a given comment. We develop two sets of regular expressions. The *element* set consists of expressions to extract location information regarding session, paragraph and sentence. The *construction* set includes patterns that express introduction, thesis, and conclusion of the paper.
- **Domain word count**: intuitively, localized comments tend to use words from a paper's topic domain. Examples of domain words extracted from History writings in our data are: *'rights', 'states', 'political', 'democracy', 'government', 'constitution'*. To build a domain vocabulary, the training data was pre-processed to extract bigrams with frequency–inverse document frequency (TF-IDF) above average. The unigrams constituting these bigrams were then considered to be the

domain vocabulary. The feature itself counts the number of domain unigrams in the comment.

– **Overlapping window size**: an overlapping-window algorithm (Ernst-Gerlach and Crane 2008) was used to search for common text spans between a feedback comment and the paper being reviewed. The algorithm iteratively searches through the paper for the referred windows of the most likely text span in the comment, and merges any two windows that are found to overlap. A larger merged window suggests more overlapped textual content, so we consider the length of the maximal window as one of our localization features. While the quoted word count feature is expected to recognize the type-3 localization (quotation excerpts, see our definition of localization types), the overlapping window size feature targets to the type-2 localization (reference to content/topic), which involves original terms but might be used in different sentence structures.

To further model the linguistic signals of localization in feedback comments (to both increase quantitative model predictive performance, and to better identify the localization text spans within comments that will be highlighted in the system's feedback), we introduce 5 additional features that can be characterized into two types:

– **Location phrase features**: 4 features fall under this category. The basic idea is to mine words and phrases that are good signals of positional localization in a semi-supervised manner. While this feature set has a similar function as the regular expression feature, these features are based on a data-driven approach to increase coverage compared to the pre-defined list of regular expressions.
– **Similarity score sum feature**: this feature supplements the domain word and window size features described above (which are based on exact lexical matching) by incorporating ideas from the Computational Linguistics literature on detecting paraphrases. While the idea of making use of lexical similarity to reason about semantic similarity at higher levels, e.g., sentence and paragraph, are very popular (Corley and Mihalcea 2005; Li et al. 2006; Islam and Inkpen 2008), we instead form different abstractions of the original sentences and apply different distance metrics to measure the similarity of every pair of abstracted strings (Malakasiotis 2009) to keep our deployment simple.

Our set of location phrase features are created as follows. Using a development set of peer feedback comments from a Cognitive Psychology class in 2007, we first collected a list of 14 location seeds which appeared to be good lexical localization signals and had occurrence greater than or equal to 50:

citations conclusion conclusions introduction page paragraph questions quotes section sentences statement statements summary transition

For each seed, we then found all words in the development data that occurred in the same context as the seed. Two words have the same context if they have the same preceding and following tokens in the corpus, e.g., *first paragraph of* and *first part of*. Next, for each word in the set of location seeds and their same-context words, we considered a bigram of the word and its preceding word to be a location bigram if the bigram indicates localization while the preceding word alone is a general term that

can be used in not-localized comment as well. The following location bigrams are used in our model:

Signal word	Preceding words
source(s)	*cite*
statement	*thesis*
paper	*begin, end*
page	*first, second, third, paragraph, on*
paragraph	*first, second, third, last, transition, next, this, full, one, two, end, sentence, open*

Finally, given the above, the location phrase feature set for each comment included the following 4 features: **number of element patterns**, **number of construction patterns**, **number of location seeds** and **number of location bigrams**. Element and construction patterns are based on the pre-defined regular expressions of the regular expression tag feature. We separate them here to better model different subtypes of positional localization. Feedback comment and signal words are stemmed before being counted.

With respect to the similarity feature, given the maximal window in the paper under review returned by the overlapping-window algorithm, we also collect two preceding and two following sentences of the overlapping window for a maximal total of 5. We generate all possible pairs of comment sentence and one of the 5 sentences found. For each pair $\mathsf{P} = \langle S_c, S_w \rangle$, in which S_c and S_w are the comment sentence and a paper sentence found by maximal window algorithm respectively, we extract from the longer sentence all subsequences seq of consecutive tokens that have the same number of tokens as the shorter sentence (pivot sentence p). We define different similarity functions that apply to pairs of strings, and given a similarity function f the similarity score of $\mathsf{P} = \langle S_c, S_w \rangle$ is aggregated as follows:

$$Sim^f(\mathsf{P}) = \max\left(Sim(seq, p)\right)$$

Then for each comment C, we calculate its maximum similarity score:

$$Sim^f(\mathsf{C}) = \max(Sim^f(\mathsf{P})) \qquad \forall \mathsf{P}$$

We use six similarity metrics:

1. Levenshtein similarity: inverse of Levenshtein distance which is the minimum number of edits (insertion, deletion, substitution) to transform a sequence to the other normalized by sequence's length.
2. Hamming similarity: number of positions at which corresponding elements are the same, normalized by sequence's length.
3. Variance similarity: total occurrences of common elements of the two sequences, normalized by total length of two sequences.
4. Trigram similarity: variance similarity of two 3-gram sequences.
5. Binary similarity: number of common elements normalized by sequence's length, element repetition is not counted as in variance similarity.
6. Cosine similarity: each input sequence is transformed to a frequency vector whose dimensions correspond to elements of two sequences, and value at each

dimension is the total occurrence of corresponding element in two sequences. Cosine similarity is cosine of the angle between two vectors calculated using Euclidean dot product formula.

We then apply them to four different abstractions of the original sentences:

1. Sequence of original tokens. For example, {*The, introduction, does, not, have, a, clear, thesis, statement, .*}
2. Sequence of part-of-speech tags, e.g., {*DT, NN, VBZ, RB, VB, DT, JJ, NN, NN, .*}
3. Sequence of tokens that are recognized as noun, i.e., have part-of-speech tag in {*NN, NNS, NNP, NNPS*}
4. Sequence of tokens that are recognized as verb, i.e., have part-of-speech tag in {*VB, VBD, VBG, VBN, VBP, VBZ*}

Overall we thus have 24 similarity functions f. Given a feedback comment, its **similarity score sum feature** is the sum of its 24 similarity score $\sum_f Sim^f(\mathsf{C})$. While pivot sentence p and the token sequence seq have the same number of tokens, their sequences of nouns (or verb) may have different length. In this case, we take the shorter as pivot sequence, and extract all sub-sequences of the longer in a similar way of what we do with $\mathsf{P} = \langle S_c, S_w \rangle$.

Pre-trained Models and Cost-Sensitive Machine Learning

The models for predicting feedback localization were pre-trained using the above features with the Logistic Regression algorithm implemented in Weka (Hall et al. 2009). We experimented with different learning algorithms, e.g., Decision Tree, Support Vector Machine, and observed that they yielded lower 10-fold cross validation performance than Logistic Regression. Since the effectiveness of supervised Machine Learning depends not only on the features available for prediction, but also on both the amount of training data and the similarity of training and testing data, we supplemented the training data used in our initial laboratory study with additional annotated data.

Because the 2013 deployment was going to be evaluated in a college class, we added an annotated corpus of peer feedback comments from a college Computer Science course and used these in addition to the original college History data (Nelson and Schunn 2009; Xiong and Litman 2010). In the 2014 deployment, the peer-review system was going to be evaluated in both college and high-school classes, so we added an annotated corpus of peer feedback comments collected from a high-school Literature course to the training corpus from the first deployment and retrained the model. Since we did not have access to new data from the actual academic disciplines that would be the testbeds for our deployments, we could only increase similarity in terms of whether the added training data came from college or high school classrooms. Label distributions of training data for the two deployments are shown in Table 3.

In addition, we biased the Machine Learning process to yield a localization prediction that would trigger the system feedback with high precision (rather than optimizing for feedback localization prediction performance). By system feedback

Table 3 Training data for feedback localization prediction

	2013 deployment	2014 deployment
LOCALIZED	1089	2063
NOT-LOCALIZED	1060	1621
Total comments	2149	3684

precision, we mean that feedback submissions which triggered system feedback are those actually in need of localization revision, i.e., have at least a feedback comment that is not localized. We decided that system feedback precision was more important than recall, because we thought it would be better to miss some feedback opportunities than to provide incorrect feedback (e.g., by telling a student to revise his/her feedback when all of the comments were already localized).

Designing the system feedback strategy, we first chose the localization threshold to trigger system feedback, i.e., a peer feedback submission would trigger system feedback if its ratio of localized comments over its total number of comments was less than the threshold value. While we aimed for detecting peer feedback submissions with at least one not-localized comment, our feedback localization model was not perfect so setting the threshold to 1 triggered almost all feedback submissions, which was not desired. Instead, we conducted a pilot study where for each feedback comment, we collected predicted labels by the feedback localization prediction model and true labels annotated by experts. By varying the threshold value, we observed that our feedback localization model best agreed with human experts with respect to triggering system feedback when the threshold was set to 0.5. Therefore, we fixed a threshold of 0.5 for all later deployments (i.e., testing on totally new data separate from the pilot study data).

Next, we implemented cost sensitive learning for our feedback localization prediction model, i.e., weighing certain types of prediction errors more heavily than others during model training using a cost-matrix. In particular, to optimize precision when triggering system feedback, we penalized predicting Localized comments as Not-localized during model training. Because our feedback localization models were trained to predict a feedback comment as Localized or Not-localized, a cost-sensitive learning paradigm required two misclassification costs as additional parameters:

- False-Negative (FN) cost: when a Localized comment was predicted as Not-localized
- False-Positive (FP) cost: when a Not-localized comment was predicted as Localized

In each deployment, best values of the misclassification costs were selected through cross-validation using training data. In the 2013 deployment, we obtained FN-cost = 5, FP-cost = 1. In the 2014 deployment, we had FN-cost = 3, FP-cost = 1. Because we did not optimize for prediction performance at the comment level, we obtained lower training performance with cost-sensitive learning.

Results and Discussion

Feedback Localization Prediction Performance

At the comment-level, we evaluate how well the feedback localization prediction models in the two deployments predicted the presence or absence of localization compared to the human annotations. We also compare the models' performance to two corresponding baselines which are Majority-class and Bag-of-words. The Majority-class model assigns all comments to the label of the majority class which is NOT-LOCALIZED for all comment sets as shown in Table 2. This is expected as most of the comments in the corpus subset used for our current annotation were from feedback submissions triggering system feedback, which should have low localization ratios. This does not imply that not-localized comments are the majority class of our comment data. In fact, in our earlier work which focused on training prediction models (as opposed to the current study, which focuses on the system feedback strategy), all of the comments in each training corpus were annotated. In those corpora we found that both localized and not-localized can be the majority class, with the majority ratios varying from as low as 52 % to as high as 72 %.

Bag-of-words (BoW) is a sparse vector model of occurrence counts of tokens in comments. We include unigrams, bigrams and trigrams as tokens because localization patterns can occur as a single word (e.g., *abstract*) or multiple-word expression (e.g., *first two sentences*). For each deployment, we train the BoW model with the corresponding training data that was used for our deployed model using a Logistic Regression algorithm. Because the cost-matrix was optimized for obtaining high system feedback precision using our proposed models, directly applying cost-sensitive learning to BoW models greatly degraded feedback localization prediction. Thus, we did not train BoW models with a cost-matrix. We, however, further improved BoW performance by adding ridge regularization to address the sparsity of ngram features. We did not stem words or remove stop words as these pre-processing reduced training performance. Actually, stemming and stopword removal also decreased overall test performance of BoW on the deployment data.

Table 4 reports the prediction performance of our pre-trained feedback localization models on peer feedback data in our two deployments. Since two of our datasets have skewed class distribution (a large majority of NOT-LOCALIZED comments), we report macro-average F1-score. The results show that our feedback localization prediction models consistently outperformed the baseline models in all four feedback comment sets with respect to all performance metrics. The F1 of NOT-LOCALIZED is higher than that of LOCALIZED, which reflects the cost-matrix used when training the models (i.e., the matrix was designed to yield high precision of system feedback) as described above.

In addition, when comparing these results with our first reported results for localization prediction (Kappa of 0.55 for the feedback localization model in Xiong and Litman (2010), where the model was cross-validated using a single dataset of peer reviews from a college history class), we see that the worst performance on our

Table 4 Localization prediction performance at the comment level of our deployed models compared to two baseline models: Majority-class (Majority) and Bag-of-words (BoW)

	Majority	BoW	Deployed	Majority	BoW	Deployed
	RM			CogPsy		
Accuracy	0.52	0.61	**0.73**	0.89	0.88	**0.92**
Kappa	0	0.20	**0.46**	0	0.49	**0.69**
F1	0.34	0.55	**0.73**	0.47	0.74	**0.84**
F1:LOCALIZED	0	0.40	**0.68**	0	0.56	**0.73**
F1:NOT	0.68	0.71	**0.77**	0.94	0.93	**0.95**
	Math1			Math2		
Accuracy	0.53	0.64	**0.65**	0.77	0.74	**0.77**
Kappa	0	0.26	**0.28**	0	0.11	**0.20**
F1	0.35	0.60	**0.61**	0.43	0.54	**0.57**
F1:LOCALIZED	0	0.47	**0.48**	0	0.24	**0.28**
F1:NOT	0.69	0.73	**0.74**	0.86	0.84	**0.86**

Performance metrics include Accuracy, Kappa, F-measure of LOCALIZED (F1:LOCALIZED) and NOT-LOCALIZED (F1:NOT) classes, Macro-average F-measure (F1) of these two classes. Higher values are bold. RM reports the prediction model from the first deployment, while CogPys and Math report the revised model from the second deployment

two Psychology college classes is only slightly degraded in our cross-course evaluation setting in which training and test data are from different courses/classes. That is, the model trained on data from college History and Computer Science classes yielded a Kappa of 0.46 when deployed in Research Methods lab. This evaluation setting is more difficult than cross-validation because the feedback localization model was trained using peer feedback comments on student papers of different writing topics, and from different academic disciplines than the test corpus of interest in this study. The best test performance was achieved during the second deployment for the Cognitive Psychology class, with cross-course Kappa of 0.69 which is even higher than the reported Kappa in the original publication (Xiong and Litman 2010). These results show that our feedback localization models obtained high prediction performance at the comment level when deployed in college classes.

We see lower prediction performance when the model that was deployed in Cognitive Psychology was later tested on the two comment sets from the Interactive Mathematics class. However, the fact that our model yielded higher performance than the Bag-of-words baseline has shown the advantages of our proposed features. The features were designed to capture linguistic styles (e.g., location phrases) and to abstract over content mentioned in the paper (e.g., domain word count, maximal window size, similarity score sum). Thus, they are expected to be more domain-adaptable than generic n-grams. In fact, out of 11 features that we used, 4 location phrase features, domain word count, and regular expression tag have the largest weights returned by Logistic Regression algorithm, which shows that they are the most effective features.

Although both high-school and college data were used to train the peer feedback localization prediction model for the 2014 deployment, results revealed difficulties that our model faced when tested on high-school data. While an analysis that investigates style and quality differences of textual comments between college and high-school students, which might cause the performance disparity among data sets, is beyond the scope of our current study, the lower performance with high-school data points to the need to either re-train the model with even larger and better data in future work, and/or develop new features better tailored to peer feedback comments in high school classes.

System Feedback Precision

At the feedback-submission level, we evaluate how often system feedback was triggered precisely, i.e., the submission that triggered system feedback was not already fully localized. Recall that in both two deployments, student reviewers were encouraged to add location information into their comments (see the formative feedback messages in Figs. 1 and 2), but they did not know the localization ratio that activated the system feedback. Thus, we consider a trigger to system feedback to be precise (with respect to students' opportunity to revise a non-localized comment) when at least one of the comments in the peer feedback submission is human-coded as NOT-LOCALIZED. We use human-annotated localization labels to compute the true localization ratio of every peer feedback submission that triggered system feedback, and label a trigger to system feedback as Incorrect if the corresponding peer feedback submission had true localization ratio of 1, and Correct otherwise. As shown in Table 5, the peer feedback localization prediction models yielded no incorrect trigger to system feedback in the Research Methods Lab and Cognitive Psychology classes. However, up to 23 % of triggers to system feedbacks were incorrect in the Interactive Mathematics class, which likely reflects the lower prediction performance at the comment level.

In sum, our results show that with real college classroom settings, our models predicted localization at the comment level accurately enough to in turn trigger system feedback with high precision. While the prediction performance with high-school peer feedback data was limited at the comment level, the system feedback precision was nonetheless promising with 77 % to 83 % of system feedback triggered correctly.

As we described, the annotated data for this study were comments of peer feedback submissions that either triggered system feedback or were revised (without immediately prior system feedback). We did not code any submissions that passed

Table 5 System feedback precision

	RM	CogPsy	Math1	Math2
Total system feedback	43	15	58	74
True-ratio < 1	43	15	48	57
True-ratio = 1	0	0	10	17
System feedback precision	1.0	1.0	0.83	0.77

Only first submissions are counted

the localization check (i.e., did not trigger the system feedback) and had no revision. Therefore, we obviously missed the False-Negative instances where the peer-review system accepted submissions that should have triggered system feedback. Due to time and resource limitations for the annotation work, and our focus on system feedback precision over recall, our current study evaluates the peer-review system on how well the system performed when it provided feedback on peer feedback submissions, which we thought more urgent than the dual evaluation on how accurately the system passed peer feedback submissions. A comprehensive analysis that covers all peer feedback submissions is left for future work.

Study 2: Reviewer Response to System Feedback

Introduction

In this study, we investigate whether student reviewers actually revised their comments in response to the system's feedback, to look for answers to Research Question 2. For the evaluation, we consider only first peer feedback submissions that triggered system feedback (see Table 2 for number of such peer feedback submissions). For each of such peer feedback submissions, we identify which button the reviewer clicked, and whether he/she edited any comments. In addition, as comments of those submissions were manually annotated for the presence of localization, we count number of comments annotated as Localized. Then, the true localization ratio of each peer feedback submission is defined as a ratio of comments annotated as Localized over total number of comments.

Method

A student reviewer can respond to the system's feedback by choosing one of three buttons as shown in Figs. 1 and 2: REVISE, EXAMPLE and DISAGREE. We, however, classify reviewer responses not exactly the same as the buttons clicked. First, clicking the EXAMPLE button shows the reviewer examples of localized comments, but does not submit the feedback. Instead the reviewer has to go back to the system feedback interface and choose one of the other two buttons to submit the feedback. Thus we do not consider viewing examples as a reviewer response in this study. In fact, our log data revealed a low number of EXAMPLE clicks. Given 190 peer feedback submissions that triggered system feedback (see Table 2), there were only 3 times that the reviewers clicked the EXAMPLE button.

Second, our formative feedback strategy did not force any constraint between the clicked button and whether any peer feedback comments were changed. That is, a student reviewer can change one or more comments then click the DISAGREE button telling the system that the "original comments" do not have any localization issues so do not need to be fixed. Conversely, a reviewer may click the REVISE button telling the system that a "revision" is made, without actually editing any of his/her comments. Examining the peer feedback data, we find no comment that is edited before the reviewer clicked DISAGREE. We however observe a large number of unmodified

comments associated with REVISE clicks. Therefore, we classify reviewer responses into three types as follows:

- REVISE: the reviewer resubmits feedback by clicking the REVISE button after revising it. This response is associated with some actual change(s) made to the feedback comments.
- 0-REVISE: the reviewer clicks the REVISE button but does not change any feedback comment. This response type is more like a disagreement but we consider it as a separate type in this study because its popularity may give us lessons of interface design and student behavior.
- DISAGREE: the reviewer disagrees with the system feedback by clicking the DISAGREE button to submit his/her feedback without revision.

Because a reviewer may open a submitted feedback and edit its comments, he/she may receive system feedback multiple times in a sequence of peer feedback submissions for the same paper. An analysis shows that Cognitive Psychology has only one system feedback triggered by a resubmission. Research Methods has 84 % of system feedbacks triggered by first peer feedback submissions. In Interactive Mathematics, system feedback on first peer feedback submissions account for 67 % and 86 % of the total system feedbacks in the first and second assignments, respectively. Because the majority of system feedback happens at the first peer feedback submission, for the present study, we do not consider reviewer responses after the system's feedback on resubmissions of peer feedback.

Results and Discussion

Table 6 shows the counts and percentages of different types of student reviewer response to system feedback. No matter whether the feedback localization prediction performance was high (in Research Methods lab and Cognitive Psychology classes) or lower (in Interactive Mathematics class, recall Table 4 and Table 5), student reviewers revised their comments (#REVISE) less than they disagreed with the system (#DISAGREE). Furthermore, in the high-school class, the number of times when student reviewers clicked REVISE but did not change any comments (#0-REVISE) is even larger than the number of actual revisions. Despite the improvement we had made to the formative feedback interface in the 2014 deployment, we do not

Table 6 Counts (cnt.) and percentages (pct.) of different types of reviewer responses to system feedback on first peer feedback submissions

	RM		CogPsy		Math1		Math2	
Response	Cnt.	Pct.	Cnt.	Pct.	Cnt.	Pct.	Cnt.	Pct.
REVISE	12	28 %	1	7 %	14	24 %	18	24 %
0-REVISE	1	2 %	1	7 %	22	38 %	33	45 %
DISAGREE	30	70 %	13	86 %	22	38 %	23	31 %
Total	43	100 %	15	100 %	58	100 %	74	100 %

see a consistent increase of #REVISE or decrease of #DISAGREE as expected across deployments.

To investigate whether student reviewers disagreed with the system feedback for good reasons (e.g., while not perfect, their feedback submissions were already highly localized), Fig. 3 reports the percentage of responses of different types with respect to different bins of true localization ratios. While a Pearson correlation test may be possible for continuous data with small size given prior knowledge of data distribution (Bland and Altman 2009), we have no assumption of distribution shape of reviewer response percentage by localization ratio. Therefore, we conduct Spearman's rank correlation tests between the percentage of responses (scaled to [0,1]) and the true localization ratio bins (10 samples) for a combination of all peer feedback data. Test results show no significant rank correlations with true localization ratio for any type of responses. Because of small numbers of reviewer responses and possible values of true localization ratio, we do not use more than 10 bins, and thus correlation test may be noisy. Nevertheless, as shown in the figure, student reviewers' disagreement and unchanged revision do not seem to relate to how well the original reviews were localized. Reviewer response percentages for individual data sets are shown in Appendix C.

While we do not have a direct evaluation of the formative feedback interface improvement (with respect to numbers of REVISE and DISAGREE) from the 2013 to the 2014 deployments, the analyses of reviewer responses suggests that peer feedback localization recall on the system feedback interface does not have much impact on student reviewers. A user study will be necessary to better understand reasons of large #0-REVISE and #DISAGREE. Moreover, our peer-review system has recently been provided a new function that checks whether a reviewer has changed any feedback comments. Based on the check, the reviewer will be guided to click the appropriate button that is REVISE if some comment(s) were edited or DISAGREE otherwise. Therefore, we expect that the recorded responses will be a more accurate representation of reviewers' intent.

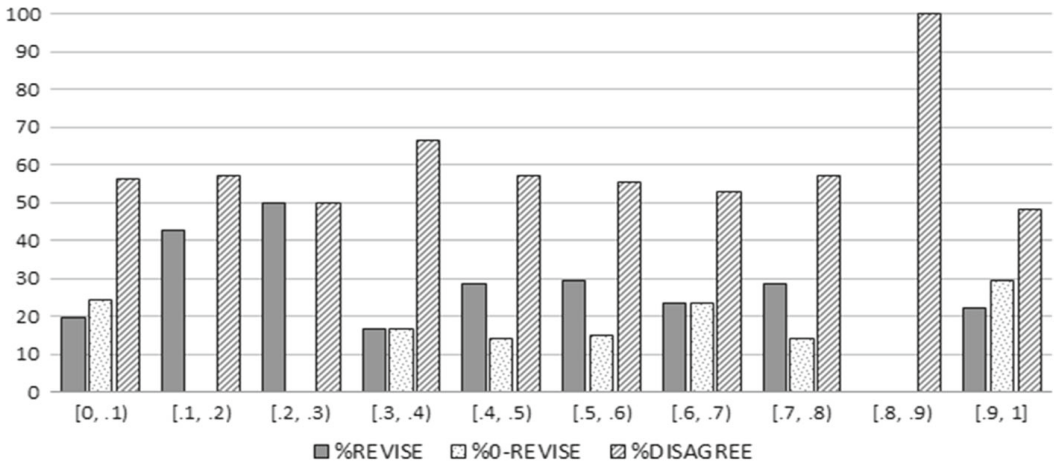

Fig. 3 Histogram of reviewer response percentage by true localization ratio. True localization ratios are placed in 10 bins. Reviewer responses of each type are summed over four peer feedback data sets

Study 3: Impact of System Feedback on Peer Feedback Revision

Introduction

This study focuses on student reviewers who indeed revised their peer feedback, to shed light on Research Question 3. We investigate whether the number of localized comments in fact increases after peer feedback revision, and whether revision behavior varies depending on the presence of system feedback. As introduced in Study 1, data for this study consists of comment pairs in which each pair includes the original comment and its revision.

Method

We evaluate the effectiveness of the system feedback by looking at the human-coded localization labels of edited comments of different edit patterns. As defined before, a reviewing session starts when a student reviewer opens the reviewing interface and ends when he/she submits the feedback successfully or closes the interface. In our 2013 deployment, a student reviewer only successfully submitted his/her feedback by either passing the feedback localization check or disagreeing with the system (by clicking on the DISAGREE button in Fig. 1). Thus, a reviewing session might contain more than one resubmission. In the 2014 deployment, if the first peer feedback submission triggered system feedback, the resubmission would not be checked for feedback localization. Thus, a reviewer would need at most two attempts to submit his/her peer feedback and end the session. By design, in both deployments student reviewers can open their previously submitted feedback at any time later, which would create new reviewing sessions. We allowed this because student reviewers may get new feedback ideas after reviewing other papers, so the system gives them opportunities to revise their current peer feedback with ideas that they have learned from their other peer feedback.

System Feedback Scopes

With an emphasis on the impact of system feedback, we consider two scopes of system feedback with respect to comment edits during a reviewing session:

- SCOPE=IN: a reviewer receives system feedback in a reviewing session, and we consider the first peer feedback resubmission that followed the system feedback. With SCOPE=IN we aim to study comment edits that occur immediately after system feedback, and thus we have evidence of a direct impact of the system feedback. We relax the condition of peer feedback resubmission so that we consider not only the immediately following resubmission in the reviewing session, but also the resubmission of the first following reviewing session. This relaxation is made to cover the cases that reviewers first disagree with the system feedback but then open their submitted feedback for revision right after.
- SCOPE=OUT: a reviewer has never received a system feedback when submitting peer feedback on the current paper, but encountered system feedback during

a prior peer feedback on a different paper. While SCOPE=IN requires that peer feedback revision must be on the same paper with the first peer feedback submission that triggered system feedback, SCOPE=OUT requires that peer feedback revision must be on a paper whose prior peer feedback submissions never triggered system feedback. By enforcing this condition, we expect to have evidence of indirect impact and retention of the system feedback (from a prior peer feedback on a different paper).

To illustrate the idea of system feedback scope, we use a feedback timeline to demonstrate peer feedback submissions in time order as shown in Fig. 4. We set T_4 when the student was reviewing paper C as the time of interest. At prior times, the student reviewed papers A, B and C, and no other paper was reviewed by the student. Thus the paper C had been reviewed twice and we have a peer feedback revision (at T_4). Now, we consider different cases of system feedback occurrences and explain how we calculate system feedback scope for the revised peer feedback at T_4.

Case 1: a system feedback occurred at time T_3 at which the student had the first submission for her feedback on paper C. In this case, scope of the system feedback at T_4 is IN because T_4 is the first peer feedback resubmission following the system feedback. The peer feedback resubmission T_4 can be of the same or a separated reviewing session with the first peer feedback submission T_3 depending on whether the reviewer clicked REVISE to submit the revision or DISAGREE to confirm the first submission. If the reviewer clicked DISAGREE then T_4 should be a new reviewing session.

Case 2: a system feedback occurred at times T_1 or T_2, and the peer feedback submission T_3 of paper C passed the feedback localization check. In this case we have system feedback scope OUT for the peer feedback resubmission T_4, even when T_4 triggered another system feedback. This later triggering, however, would have the scope with respect to later resubmissions of peer feedback on paper C if any.

Case 3: none of submissions T_1, T_2 or T_3 trigger a system feedback. This case is not considered in our study because we hypothesize that peer feedback revision at T_4 may be due to some other reason than peer feedback localization.

Comment Edit Patterns

For each value of system feedback scope, we collect all peer feedback comments that were edited in peer feedback revisions and compare each comment's true localization

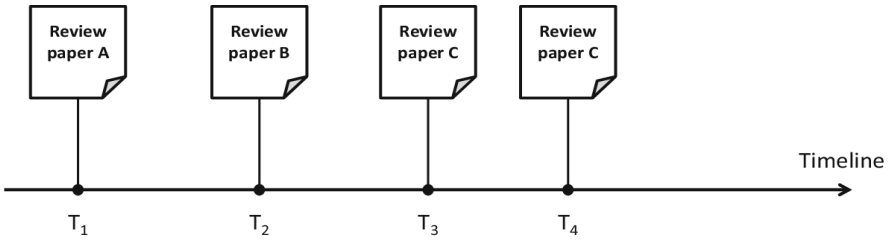

Fig. 4 Feedback timeline of a student reviewer found in our peer feedback data. Consider time T_4 at which the student was reviewing the paper C, and assume that before T_4, the student reviewed papers A, B and C at times T_1, T_2 and T_3 respectively. We have a feedback revision when the paper C had been reviewed twice

label to the true label of its previous version. The edit pattern of most interest is NOT-LOCALIZED → LOCALIZED, as this is the type of successful edit that the formative feedback strategy was designed to promote. At the other extreme, the least desirable pattern is LOCALIZED → NOT-LOCALIZED, as this type of comment edit decreases feedback quality with respect to localization. In different contexts, the patterns LOCALIZED → LOCALIZED and NOT-LOCALIZED → NOT-LOCALIZED may reveal different meanings which we will discuss in detail shortly.

Results and Discussion

Table 7 reports numbers of comment pairs in four peer feedback data sets according to the four comment edit patterns with respect to localization label. First we observe that the total number of comment edits in the SCOPE=IN condition (86) is much larger than that of the SCOPE=OUT condition (15) for all four peer feedback data sets. This reveals that a direct system feedback showed a stronger impact than an indirect one.

Next we compare different edit patterns in Table 7 across peer feedback data sets. We count the total number of each edit pattern:

Edit patterns	SCOPE=IN	SCOPE=OUT	Total comments
NOT-LOC.→LOCALIZED	27	8	35
LOCALIZED→LOCALIZED	26	3	29
NOT-LOC.→NOT-LOC.	33	4	37
LOCALIZED→NOT-LOC.	0	0	0

Table 7 Comment edit patterns by system feedback scopes

	SCOPE=IN		SCOPE=OUT		SCOPE=IN		SCOPE=OUT	
	RM				CogPsy			
Total reviews	12		4		2		1	
NOT-LOC.→LOCALIZED	8	20 %	2	50 %	0	0 %	1	100 %
LOCALIZED→LOCALIZED	13	33 %	1	25 %	0	0 %	0	0 %
NOT-LOC.→NOT-LOC.	19	47 %	1	25 %	2	100 %	0	0 %
LOCALIZED→NOT-LOC.	0	0 %	0	0 %	0	0 %	0	0 %
Total comments	40	100 %	4	100 %	2	100 %	1	100 %
	Math1				Math2			
Total reviews	17		5		20		3	
NOT-LOC.→LOCALIZED	8	34 %	3	43 %	11	55 %	2	67 %
LOCALIZED→LOCALIZED	9	38 %	1	14 %	4	20 %	1	33 %
NOT-LOC.→NOT-LOC.	7	30 %	3	43 %	5	25 %	0	0 %
LOCALIZED→NOT-LOC.	0	0 %	0	0 %	0	0 %	0	0 %
Total comments	24	100 %	7	100 %	20	100 %	3	100 %

The number of peer feedback submissions that have system feedback with SCOPE=IN can be larger than the number of REVISE responses because of our relaxed condition of peer feedback resubmissions as described

We observe that in a slight majority of opportunities, student reviewers failed to add location information to not-localized comments (35 NOT-LOCALIZED → LOCALIZED versus 37 NOT-LOCALIZED → NOT-LOCALIZED). Thus, we do not have evidence supporting that student reviewers were more likely to newly localize comments given system feedback. However, the results still show promising findings. First, the least desirable pattern of LOCALIZED → NOT-LOCALIZED did not occur in any classes of the two deployments. The lack of occurrence of this pattern suggests that our highlighting of localized comments in green (2013 deployment) and further highlighting localization text within these comments in bold (2014 deployment) might help student reviewers not to remove localization from their localized comments.

Second, the pattern NOT-LOCALIZED → LOCALIZED which corresponds to the most desirable edit contributes the second largest portion of edits over all conditions (35 %). Especially, in SCOPE=OUT reviewers gave more successful edits than other edit patterns. Thus, the number of localized comments consistently increased after system's feedback across deployment data. Such evidence indirectly suggests that the system's feedback indeed provided student reviewers an opportunity to localize their previously unlocalized comments, and the impact still remained in later reviewing sessions after the system feedback is removed.

Manually examining peer feedback comments of the edit pattern NOT-LOCALIZED → LOCALIZED, we observed that in all cases the student reviewers added the section headings or quotes to localize their comments.

- *You should include your 16 trials. → You should include your 16 trials in your **process section**.*
- *A lot of grammar problems → A lot of grammar problems in the **problem statement , process and solution***
- *There are some punctuation error and grammatical errors here and there. Make sure to proofread. → There are some punctuation error and grammatical errors here and there. Make sure to proofread. eg **"Thesuggest that diffusion of responsibility..."** This toward the **end of the discussion section**.*

We also observe three interesting cases where the reviewers changed their comments from a praise to a criticism when they added localization. As the examples in the system message encouraged localization to be associated with describing problems and providing solutions to problems in the reviewed papers, this kind of edit reflects an additional impact of system feedback. In the examples below, some student reviewers used the abbreviation "*pow*" for the phrase "*problem of the week*".

- *I think this essay is great so I'm not sure. → I think this essay is great so I'm not sure. I think you should add more to your **Extension** because it is a little hard to understand, but other than that, its fine.*
- *You have everything that is needed in this p.o.w → You need a little bit more information on how you got your **answer in your solution***
- *Very precise and replicable. Every aspect of the study is addressed → Very precise and replicable [...] Some details, like **"manila folder" and "five-foot***

radius" *may be too intricate and unnecessary, nonetheless they do not delineate from the vital information.*

We manually examined all 29 comment edits of LOCALIZED → LOCALIZED and observed that 13 of these edits (45 %) introduced more localization in the revised comments. All of the new location information added to the revisions are in the form of paper section expressions (e.g., *your problem statement*, *first paragraph*) and quotes (e.g., *"Population Density and its Effect on Willingness to Help". Above or below or instead of that heading*). The other 16 edits (55 %) added general judgments or suggestions (e.g., *Finish your table and this would be perfect*, *you need to have step by step on how you get the answer*).

Out of the 37 edits of type NOT-LOCALIZED → NOT-LOCALIZED, 15 edits (41 %) showed evidence of attempting to localize comments but failing because the added information was too general/vague to be annotated as localized (e.g., *add more in the statement and process in the middle*, *any sections*, *throughout your whole POW*). These findings of LOCALIZED → LOCALIZED and NOT-LOCALIZED → NOT-LOCALIZED patterns suggest that student reviewers did indeed have an intention to revise their comments with respect to localization given the system's feedback (both direct and indirect), but they just did not efficiently localize the comments or add more useful information, i.e., adding coarse-grain localization or too vague information.

General Discussion

Automated Formative Feedback Strategy for Increasing Peer Feedback Localization

We have presented our research in developing and evaluating a web-based peer-review system that provides interactive and formative feedback regarding peer feedback localization immediately to peer reviewers (i.e., at the time of peer feedback submission) whenever their written peer feedback comments are evaluated to be of low quality with respect to localization. We also presented the results of an extrinsic evaluation that examined both the immediate and retention effects of the system's feedback on reviewers' revision behavior, using data from all classroom deployments of our system.

Our comment-level evaluation results showed that our peer feedback localization prediction models outperformed Majority-class and Bag-of-words baselines with absolute performance levels approaching prior laboratory results (Xiong and Litman 2010) for the peer feedback data of the college classes. However, the prediction performance was weaker for the high-school class. Nonetheless, our feedback submission-level results demonstrated that with an optimization to favor precision over recall in triggering system feedback, the peer feedback localization model for high-school data could obtain high system feedback precision ranging from 77 % to 83 %, despite the limited prediction performance at the comment-level.

Analyzing reviewer responses to the system's feedback, we observed a large number of student disagreements with the system's suggestion to increase localization, as well as a large number of self-reported peer feedback revisions that actually contained no edits. Despite the changes we had made to the formative feedback interface to better call reviewers' attention to peer feedback localization, the number of disagreements and unchanged revisions remained substantial in both deployments. Moreover, these two types of responses seemed not to relate to the actual (gold-standard) percentage of localization in the peer feedback.

In addition, we found that for reviewers who revised their peer feedback after the system's feedback, the number of comments with localization increased after editing. Moreover, the system feedback appeared to improve localization even in later reviewing sessions that did not trigger system feedback. This makes us hope that interactive and formative system feedback such as proposed here may facilitate robust learning and improve student's feedback localization skills even after the system feedback is removed. However, the results also demonstrated that our current approach could be further improved, as there were a large number of unsuccessful attempts to localize comments when edits were indeed made after the system feedback.

Limitations

As discussed in Study 1, a major limitation of our present research was that we annotated only feedback comments of peer feedback that triggered the system's feedback. Thus, we do not have any evidence regarding how likely our formative feedback strategy accepts the peer feedback which should have instead triggered system feedback. If a future study annotates comments from all peer feedback, we can optimize a peer feedback localization prediction model for system feedback's F1 score rather than precision, and explore the educational impact of doing so. Consequently, the formative feedback strategy can be aimed at targeting peer reviewers that are in need of different types of help for feedback localization.

Furthermore, our formative feedback strategy was not deployed in a large scale, which in turn did not give us enough data to test for significance of our findings in this research. First, due to a small number of student reviewer responses to the system's feedback, we cannot conduct correlation tests for a relation between peer feedback localization ratio and likelihood of student response (Study 2). Second, while student reviewers were more likely to add location information to their non-localized comments when system feedback scope was OUT, the number of comment edits was too small to conclude that such a finding was significant and consistent across classes. We hope that more deployment data will show the lack of a relation between student reviewer disagreement with system feedback and the peer feedback localization ratio, as well as demonstrate that our formative feedback strategy has significant long-term impact on increasing peer feedback localization when system feedback does not later occur.

Implications for Practice and Future Work

We believe that the results regarding the role of interactive and formative system feedback for peer feedback localization could encourage research on improving other measures of peer feedback quality such as feedback solution and justification, especially given recent work on prediction models for a variety of peer feedback characteristics, e.g., Xiong et al. (2012), Nguyen and Litman (2014), Ramachandran and Gehringer (2015), and Ramachandran et al. (2016). Moreover, the experimental results of indirect impact of system's feedback make it worth comparing and combining interactive feedback with summative feedback.

A notable finding from our research is that student reviewers seemed to disagree with system feedback no matter how well their peer feedback comments were localized. It suggests that student responses to system feedback may not directly be affected by the localization ratio of their peer feedback, but by other factors currently not accounted for in our present research. In an effort to better help student reviewers improve their reviews, educators and researchers need to improve both feedback characteristic prediction performance as well as formative feedback methods to better direct student reviewers to revise their comments.

One of the next steps for our research is to conduct a larger deployment of the automated formative feedback strategy for peer feedback localization, and to annotate comments from peer feedback that both do and do not trigger system feedback. The main goal is to address the limitations of our current studies and look for stronger evidence of the impact of our formative system feedback on peer feedback localization improvement. However, there are also several other possible future directions worth investigating.

We plan to improve both the feedback methods used in the system feedback interface and the feedback localization prediction model to provide more accurate prediction and better help students localize their feedback comments when they do receive feedback from the system. For example, depending on the feedback prompt, not all feedback comments can or should be localized, so we would like to incorporate feedback prompt analysis into feedback localization prediction. When the comments have associated numeric ratings, this information might also be incorporated into localization prediction. Our current analysis makes the assumption that the training instances are conditionally independent given the features. Removing this assumption is another area for our future investigation (Goldin 2012; Piech et al. 2013; Waters et al. 2015). We also plan to do further annotation to examine not only whether, but how strongly, a comment is localized. This will provide us with a more nuanced way to quantify student learning with respect to writing localized comments.

In addition, we plan to interview student reviewers about why they were disagreeing with the system feedback, as our initial analyses did not show any relationship with true localization. Also, our analysis of system feedback scope needs to be followed up with a controlled study, to demonstrate how much of the unprompted localization edit was due to retention of system feedback.

Finally, we are currently extending our approach to provide formative feedback to reviewers on other measures of peer feedback quality beyond localization. In particular, we have developed a strategy for providing formative feedback regarding the presence of solutions in peer feedback comments that provide a critique of a paper. That formative feedback strategy had a pilot deployment in a new classroom with high-school students (Nguyen et al. 2016). Incorporating localization and solution feedback into a single system will also allow us to investigate whether providing feedback on peer feedback localization will have indirect benefits such as leading a reviewer to also improve his/her peer feedback with respect to solutions, and vice-versa.

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 NSF 1122504. We are grateful to our SWoRD colleagues for sharing the data and their feedback on our research.

Appendix A: Peer Feedback Prompts in Different Classes

Research Methods 2013

The Research Methods Lab is the laboratory component of an undergraduate course in Psychology. Students in the lab were asked to conduct an observational study that included a literature search, data collection and analysis, and a written report. Every student report was submitted to SWoRD for feedback from at most 4 peer reviewers. Each report was required to be written in an APA paper style and was reviewed according to eight different aspects specified in an instructor-defined rubric. This appendix contains the detailed feedback prompts specified in this rubric, as well as for the rubrics for the classes and assignments discussed below.

1. Abstract (1-3 comments): - All required information included? - 150 words or less; concise, specific, and accurate? - Appropriate level of detail? Comment on ways that this section failed or succeeded at doing these things. Describe anything that was missing or weak and make specific suggestions for additions, deletions, or changes to this section of the paper.
2. Introduction (1-3 comments): - Central topic introduced and background information provided? - Brief high-level overview of study design and clear statement of hypotheses? - Appropriate integration of conflicting research findings into a convincing argument for at least one hypothesis? Comment on ways that this section failed or succeeded at doing these things. Describe anything that was missing or weak and make specific suggestions for additions, deletions, or changes to this section of the paper.
3. Method (1-3 comments): - Participants adequately and accurately described? - Procedures presented accurately and clearly so study can be replicated? - Appropriate level of detail that excludes inconsequential details? Comment on ways

that this section failed or succeeded at doing these things. Describe anything that was missing or weak and make specific suggestions for additions, deletions, or changes to this section of the paper.

4. Results (1-3 comments): - Descriptive statistics reported either in text or table/figure? (paper should include both a table and figure) - Statistical tests reported completely and accurately? - Tables/figures correctly referenced in text? - Results worded so they're clearly linked to hypotheses/research questions? Comment on ways that this section failed or succeeded at doing these things. Describe anything that was missing or weak and make specific suggestions for additions, deletions, or changes to this section of the paper.
5. Discussion (1-3 comments): - Main findings summarized? - Results clearly and accurately interpreted? - Current study put into context in relation to previous work? - Strengths/weaknesses, alternative explanations, implications, suggestions for future research discussed as needed? Comment on ways that this section failed or succeeded at doing these things. Describe anything that was missing or weak and make specific suggestions for additions, deletions, or changes to this section of the paper.
6. Global Writing (1-3 comments): - Writing clear and concise, not wordy or confusing? - Ideas well organized, part of a coherent argument, flow together well? - Tone appropriate for readership of professional psych journal?
7. Technical Writing (1-3 comments): - Sentences complete and grammatically correct? - Paper carefully proof-read and spell-checked?
8. APA Style (1-3 comments): Is APA style used correctly for the following? - Numbers - Statistics - In-text citations - Paper header - Abbreviations - Section headings Etc. Are the following elements formatted according to APA style? - Abstract - Introduction - Method - Results - Discussion - References - Table/Figure.

Cognitive Psychology 2014

The Cognitive Psychology writing assignment required students to select one of a set of journal articles made available on the course website and write an article of about 700 words that was similar to articles in The New York Times' science section (e.g., explain what the main findings are and why those findings are important). Each student paper was reviewed by at most 4 peer reviewers, and on four aspects of the writing.

1. Author's Focus (1-4 comments): at the end of the paper, the writer should have mentioned what aspect of writing they wanted to practice. Give them some comments about how well they did on this dimension and also some comments about how they could have improved further.
2. Your Focus (1-4 comments): now consider what aspects of your own writing skills you want to improve. What aspect of this paper would you change to improve this aspect of the paper (and name it in your fist comment)? If you can't think of an improvement on this aspect, talk about why you think the author was particularly successful.

3. Biggest Issues (1-3 comments): name what you think is the biggest problem or two in the current paper (what most caused the ratings to be lower), and give some suggestions for how to repair it.
4. Fundamental Writing Issues (1-5 comments): if there were some common problems at the basic writing level (spelling/grammar, poor word choice, awkward sentence structures), please describe the type of problem including the location of one instance.

Interactive Mathematics 2014

The two Interactive Mathematics writing assignments were mathematics challenges that required students to write essays in which they stated the problem, described their solutions, and proved that the solutions were correct. Each paper was reviewed by at most 3 peer reviewers. Reviewers were assigned randomly and were not necessarily the same across the two assignments. Student papers of the first and second assignments were reviewed for 2 and 1 aspects respectively.

Feedback prompts of the writing assignment #1:

1. Keystone Reasoning (1-5 comments): provide feedback on the reasoning or explanation in the paper.
2. Keystone Transition (1-5 comments): comment on how well the ideas in the paper flow logically and are linked by effective transitions.

Feedback prompt of the writing assignment #2:

1. Possible Improvements (1 comment): what suggestions would you give the writer to make this essay stronger?

Appendix B: Work-flow and Interfaces of SWoRD with Automated Formative Feedback

Peer Review Process in Original SWoRD

In phase 1, student authors create first drafts and submit to SWoRD. During phase 2, student reviewers download their set of assigned drafts to review then provide feedback on each draft using the interface for submitting peer feedback (Fig. 5). SWoRD generates set of papers for each student to review based on the instructor's policy for how many and which papers each student should review. Draft-reviewer assignment can be predefined by the instructor or randomly generated by SWoRD. In phase 3, student authors view feedback on their papers from their peer reviewers using the interface for receiving peer feedback (Fig. 8). In phase 4, student authors revise their papers, hopefully based on the feedback they received from as well as provided to their peers.

Figure 5 shows parts of a student's review of the first draft of FreddieMercury12's paper "Perception through Human Memory," The paper was submitted in phase 1 in response to the "Assignment Description." Review feedback is in

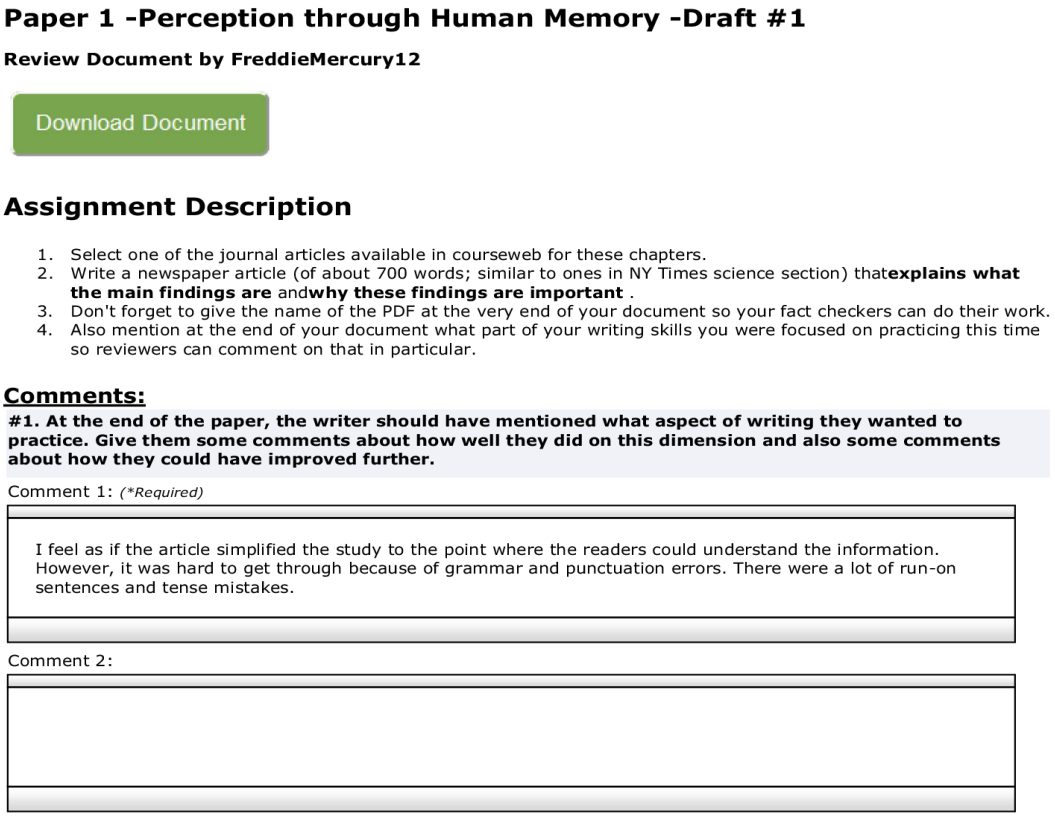

Fig. 5 Screenshot of SWoRD's interface (for reviewers) for reviewing the paper of another student. To preserve anonymity, all students are asked to create a SWoRD pseudonym, e.g., FreddieMercury12. For readability, only the first feedback prompt is shown; the numerical review rating parts are shown in Fig. 7

the form of both written comments (shown in the boxes, e.g. "I feel as if...") and numeric ratings (Fig. 7) that are associated with a set of instructor-defined rubric prompts (e.g. "#1. At the end..."). More comments of this feedback are provided in Fig. 6. The feedback rubric associated with the writing assignment (as exemplified in Figs. 5 and 6) is shown in Appendix A. The rubric specifies a minimum and maximum number of allowable comments for each feedback prompt.

Each row in the SWoRD's interface for receiving peer feedback (Fig. 8) corresponds to the "short name" of each feedback prompt that was defined in the associated rubric. Each column in the figure presents the text comments provided by one of the assigned reviewers. For example, the comments entered in Fig. 6 are shown in the column "Reviewer #1".

In addition to the four main phases, two other phases are also frequently used in SWoRD. In phase 5, student authors back-evaluate their received peer feedback comments by both rating and describing how helpful the written comments were for revising their papers (see the "Backevaluation" portion of the interface shown

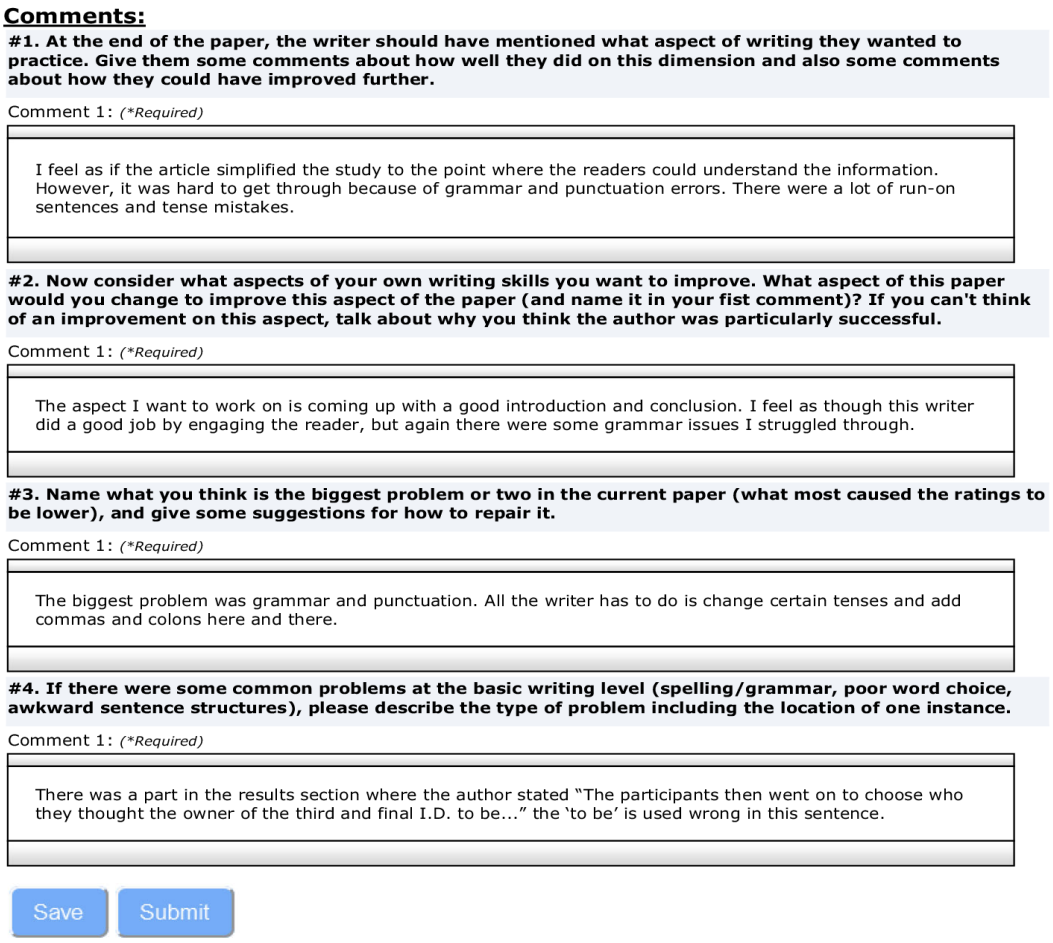

Fig. 6 Comments edited by a student reviewer using SWoRD's interface for reviewing

in Fig. 8). In a final phase, the revised paper may be resubmitted to SWoRD for another round of peer feedback. SWoRD allows the instructor to assign either the same reviewers of the prior draft or a new set of reviewers.

Automated Formative Feedback in SWoRD

Figures 1 and 2 show examples of predicted localized (part B) and not localized (part C) comments from paper reviews from two of the datasets in our study. The feedback message which suggests peer feedback revision and provides advice for doing so is pictured as in part A. When a student reviewer chooses to view model comments, examples of localized comments are displayed as in Fig. 9. The provided examples are pre-selected from the data which was used to train the peer feedback localization prediction model, thus the examples are not adaptive to the writing assignments in which the model is deployed.

Ratings:

#1. Extent to which the author described the key methods of the research article correctly.

- (•) All key method details present and accurate
- ( ) 6
- ( ) Some key method details missing, but details present are accurate
- ( ) 4
- ( ) Minor errors in presented method details
- ( ) 2
- ( ) Major errors in presented method details

#2. Extent to which the author described the key results of the research article correctly.

- (•) All key results details present and accurate
- ( ) 6
- ( ) Some key results details missing, but details present are accurate
- ( ) 4
- ( ) Minor errors in presented results details
- ( ) 2
- ( ) Major errors in presented results details

Fig. 7 Example of numerical review ratings from the review in Fig. 6.

	Reviewer #1	**Reviewer #3**
The author's focus	I feel as if the article simplified the study to the point where the readers could understand the information. However, it was hard to get through because of grammar and punctuation errors. There were a lot of run-on sentences and tense mistakes. ***Backevaluation (5)*** Yes it did!	The writing goal was met. The way that the writer explained the study was easy to understand and explained very well. It seems that although they explained the study very well, they did not further the paper by talking about what this study means for things outside of just using people's fake IDs. I think it is also important to mention that none of the people in the study figured out that the same man was on all of the IDs. ***Backevaluation (5)*** Yes!
Your focus	The aspect I want to work on is coming up with a good introduction and conclusion. I feel as though this writer did a good job by engaging the reader, but again there were some grammar issues I struggled through. ***Backevaluation (5)*** Yes	The writing skills were very clear and concise. The writer used proper punctuation, the paper flowed well, and keeps the reader interested. ***Backevaluation (3)*** No suggestion.
Biggest issues	The biggest problem was grammar and punctuation. All the writer has to do is change certain tenses and add commas and colons here and there. ***Backevaluation (5)*** Will go back and review the paper!	Give examples of how this study pertains to the outside world besides usi8ng a friend's IDs. What is the importance of the study? Maybe mention how different all of the pictures of people in the study are. They all have different hair, facial structures, skin tones. ***Backevaluation (1)*** The pictures were actually all similar to the original card owner.
Fundamental writing issues	There was a part in the results section where the author stated "The participants then went on to choose who they thought the owner of the third and final I.D. to be..." the 'to be' is used wrong in this sentence. ***Backevaluation (5)*** Will change that! Yes it helped!	Bottom of page 1, change someone to something. ***Backevaluation (5)*** Will change this and look through for more errors.

Fig. 8 SWoRD's interface (for authors) displaying all peer feedback comments on a paper. For each feedback prompt (*the rows*), the comments by different peer reviewers (*the columns*) are listed. If the student author gave back-evaluations on the received peer feedback comments, the back-ratings and back-comments are displayed under the comments. For readability, the comments of only two reviewers are shown

Instructions	You can indicate where your comments apply by: (1) Specifying page numbers and paragraph numbers in the author's text to which your comment refers, (2) Referring explicitly to the specific topic that your comment addresses, (3) Quoting the excerpt from the author's text to which your comment refers
Example #1	**On page [x] paragraph [y], ...** ***On page 2 paragraph 1**, you stated that "for every step forward, several steps back were taken". We did not go backward; we failed to move ahead as fast as we should have. You can remove it.*
Example #2	**When you talk about [topic] ...** ***When you talked about amendments that give more rights to the African Americans**, you say African Americans lost rights, but I'm not sure that they had any prior to 1865. If they did, please provide more evidence for this statement.*
Example #3	**When you write [quoted sentence from paper] ...** ***When you write "After the presidential election of 1876, and the removal of the federal troops in the south, the Compromise of 1877, ended up having little worth and resulting in the One-Party Democrats system?"**, the sentence makes no sense because it is missing s subject. What "ended up having little worth?"*

Fig. 9 Sample localization templates and model comments with localization text provided to student reviewers when they click the EXAMPLE button of the system feedback interface. Model comments were selected from a peer feedback comment corpus of a History class

Appendix C: Reviewer Response Percentage by True Localization Ratio

See Table 8.

Table 8 Histogram of responses by true localization ratios

Ratio bin	[0, .1)	[.1, .2)	[.2, .3)	[.3, .4)	[.4, .5)	[.5, .6)	[.6, .7)	[.7, .8)	[.8, .9)	x[.9, 1)	[1]
	RM										
Responses	3	4	3	4	5	7	12	5	–	–	–
%REVISE	0	50	33	25	40	14	25	40	–	–	–
%0-REVISE	0	0	0	0	0	0	8	0	–	–	–
%DISAGREE	100	50	67	75	60	86	67	60	–	–	–
	CogPsy										
Responses	11	3	–	–	1	–	–	–	–	–	–
%REVISE	0	33	–	–	0	–	–	–	–	–	–
%0-REVISE	9	0	–	–	0	–	–	–	–	–	–
%DISAGREE	91	67	–	–	100	–	–	–	–	–	–
	Math1										
Responses	16	–	1	2	1	20	5	2	1	–	10
%REVISE	19	–	100	0	0	35	20	0	0	–	20
%0-REVISE	19	–	0	50	100	20	60	50	0	–	20
%DISAGREE	62	–	0	50	0	45	20	50	100	–	60

Table 8 (continued)

Ratio bin	[0, .1)	[.1, .2)	[.2, .3)	[.3, .4)	[.4, .5)	[.5, .6)	[.6, .7)	[.7, .8)	[.8, .9)	x[.9, 1)	[1]
	Math2										
Responses	57	–	–	–	–	–	–	–	–	–	17
%REVISE	24	–	–	–	–	–	–	–	–	–	24
%0-REVISE	30	–	–	–	–	–	–	–	–	–	35
%DISAGREE	46	–	–	–	–	–	–	–	–	–	41

'-' means the bin has no data. As peer reviews of Interactive Mathematics–assignment 2 (Math2) include one comment each, the true localization ratio is either 0 or 1

References

Berg, I.vd., Admiraal, W., & Pilot, A. (2006). Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education*, *31*(3), 341–356. doi:10.1080/03075070600680836, http://tlc.zmml.uni-bremen.de/resource_files/resources/390/Design_principles_and_outcomes_of_peer_assessment_in_higher_education.pdf.

Bland, J.M., & Altman, D.G. (2009). Analysis of continuous data from small samples. *Bmj*, *338*, a3166.

Cho, K. (2008). Machine classification of peer comments in physics, In Baker, R.S.J.D., Barnes, T., & Beck, J.E. (Eds.) *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 192–196). Canada: Montreal.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328–338. doi:10.1016/j.learninstruc.2009.08.006, http://www.sciencedirect.com/science/article/pii/S0959475209000747, unravelling Peer Assessment.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology*, *103*(1), 73–84. doi:10.1037/a0021950.

Cho, K., & Schunn, C.D. (2007). Scaffolded writing and rewriting in the discipline: a web-based reciprocal peer review system. *Computers & Education*, *48*(3), 409–426. doi:10.1016/j.compedu.2005.02.004.

Cho, K., Schunn, C.D., & Kwon, K. (2007). Learning writing by reviewing in science. In *8th International conference on computer-supported collaborative learning, International Society of the Learning Sciences, New Brunswick, NJ, USA* (pp. 141–143). http://portal.acm.org/citation.cfm?id=1599600.1599626.

Corley, C., & Mihalcea, R. (2005). Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Association for Computational Linguistics, Stroudsburg, PA, USA, EMSEE '05* (pp. 13–18). http://dl.acm.org/citation.cfm?id=1631862.1631865.

Ellis, J. (2011). Peer feedback on writing: Is on-line actually better than on-paper? *Journal of Academic Language and Learning*, *5*(1), A88–A99.

Ernst-Gerlach, A., & Crane, G. (2008). Identifying quotations in reference works and primary materials. In *Research and Advanced Technology for Digital Libraries, Springer Berlin Heidelberg, Lecture Notes in Computer Science*, (Vol. 5173 pp. 78–87). doi:10.1007/978-3-540-87599-4_9.

Ferris, D.R., Liu, H., Sinha, A., & Senna, M. (2013). Written corrective feedback for individual {L2} writers. *Journal of Second Language Writing*, *22*(3), 307–329. doi:10.1016/j.jslw.2012.09.009, http://www.sciencedirect.com/science/article/pii/S1060374312000811.

Gan, M.J.S., & Hattie, J. (2014). Prompting secondary students' use of criteria, feedback specificity and feedback levels during an investigative task. *Instructional Science: An International Journal of the Learning Sciences*, *42*(6), 861–878. https://www.learntechlib.org/p/167992.

Gielen, M., & De Wever, B. (2015). Structuring the peer assessment process: a multilevel approach for the impact on product improvement and peer feedback quality. *Journal of Computer Assisted Learning*, *31*(5), 435–449. doi:10.1111/jcal.12096.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20*(4), 304–315. doi:10.1016/j.learninstruc.2009.08.007, http://www.sciencedirect.com/science/article/pii/S0959475209000759, unravelling Peer Assessment.

Goldin, I.M. (2012). Accounting for peer reviewer bias with bayesian models. In: Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems, Citeseer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18. doi:10.1145/1656274.1656278.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487, http://rer.sagepub.com/content/77/1/81.abstract.

Heift, T. (2004). Corrective feedback and learner uptake in CALL. *ReCALL*, *16*(02), 416–431. doi:10.1017/S0958344004001120.

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, *2*(2), 10:1–10:25. doi:10.1145/1376815.1376819.

Kaufman, J., & Schunn, C. (2011). Students' perceptions about peer assessment for writing: their origin and impact on revision work. *Instructional Science*, *39*(3), 387–406. doi:10.1007/s11251-010-9133-6.

Kern, V.M., Saraiva, L.M., & dos Santos Pacheco, R.C. (2003). Peer review in education: Promoting collaboration, written expression, critical thinking, and professional responsibility. *Education and Information Technologies*, *8*(1), 37–46. doi:10.1023/A:1023974224315.

Van der Kleij, F.M., Eggen, T.J.H.M., Timmers, C.F., & Veldkamp, B.P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, *58*(1), 263–272. doi:10.1016/j.compedu.2011.07.020, http://www.sciencedirect.com/science/article/pii/S0360131511001783.

Kumar, A. (2010). Error-flagging support for testing and its effect on adaptation. In *Intelligent Tutoring systems, Springer Berlin Heidelberg, Lecture Notes in Computer Science*, (Vol. 6094 pp. 359–368).

Landry, A., Jacobs, S., & Newton, G. (2014). Effective use of peer assessment in a graduate level writing assignment: a case study. *International Journal of Higher Education*, *4*(1), 38.

Li, Y., Mclean, D., Bandar, Z., O'Shea, J., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1138–1150. doi:10.1109/TKDE.2006.130.

Lippman, J., Elfenbein, M., Diabes, M., Luchau, C., Lynch, C., Ashley, K., & Schunn, C. (2012). 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*.

Lundstrom, K., & Baker, W. (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. doi:10.1016/j.jslw.2008.06.002.

Malakasiotis, P. (2009). Paraphrase recognition using machine learning to combine similarity measures. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, ACLstudent '09* (pp. 27–35). http://dl.acm.org/citation.cfm?id=1667884.1667889.

McCarthey, S., Magnifico, A.M., & Kline, S.M. (2013). Secondary students' use of two online peer review tools. Dallas, Texas.

McCarthey, S., Magnifico, A.M., & Kline, S.M. (2014). Reconsidering peer feedback for argumentative essays. Philadelphia, Pennsylvania.

Mulder, R.A., Pearce, J.M., & Baik, C. (2014). Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education*, *15*(2), 157–171. doi:10.1177/1469787414527391, http://alh.sagepub.com/content/15/2/157.abstract.

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. *Digital Education Review*, *23*, 7–26. https://www.learntechlib.org/p/131614.

Nelson, M., & Schunn, C. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, *37*(4), 375–401. doi:10.1007/s11251-008-9053-x.

Nguyen, H., & Litman, D. (2013). Identifying localization in peer reviews of argument diagrams. In *Artificial Intelligence in education, Springer Berlin Heidelberg, Lecture Notes in Computer Science*, (Vol. 7926 pp. 91–100).