eRevise: Using Natural Language Processing to Provide Formative Feedback on Text Evidence Usage in Student Writing

H. Zhang, A. Magooda, D. Litman, R. Correnti, E. Wang, L.C. Matsmura, E. Howe, R. Quintana
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, Pennsylvania 15260

Abstract

Writing a good essay typically involves students revising an initial paper draft after receiving feedback. We present eRevise, a web-based writing and revising environment that uses natural language processing features generated for rubric-based essay scoring to trigger formative feedback messages regarding students' use of evidence in response-to-text writing. By helping students understand the criteria for using text evidence during writing, eRevise empowers students to better revise their paper drafts. In a pilot deployment of eRevise in 7 classrooms spanning grades 5 and 6, the quality of text evidence usage in writing improved after students received formative feedback then engaged in paper revision.

Introduction

With benefits such as minimizing human effort and assuring scoring consistency, natural language processing (NLP) has been used to develop many Automatic Essay Scoring (AES) systems that can reliably assess the content, structure, and quality of written prose (Shermis and Burstein 2003; 2013). However, before providing students with final essay scores, engaging students in cycles of essay drafting and revising after feedback is also essential (Graham, Harris, and Santangelo 2015). This is because scoring without feedback is typically not enough to guide students on how to revise an essay for further improvement. Standalone AES systems are thus increasingly being incorporated into Automated Writing Evaluation (AWE) systems (Dikli 2006; Roscoe et al. 2014; Weigle 2013), which provide students with formative feedback in addition to (or instead of) essay scores. Formative feedback can guide students during revision in ways that help students compensate for identified essay weaknesses. Although Foltz and Rosenstein (2015) showed that student writing could improve with revisions based on AWE feedback and Chapelle, Cotos, and Lee (2015) showed that successful revising is related to feedback accuracy, much AWE research remains to be done.

This paper presents the design and first classroom evaluation of eRevise, an AWE system for improving students' ability to use text evidence – a dimension of writing that is important for college and career readiness. eRevise has been designed for students in grades 5-6 taking the Response to Text Assessment (RTA) (Correnti et al. 2013), where students first write an essay in response to a source text and are then assessed using a detailed Evidence scoring rubric.[1] In particular, eRevise has been developed for deployment in a formative/classroom environment over two class periods (in contrast to a summative/high-stakes usage). Students write their essays in the first period, then revise their essays in the second period after receiving formative feedback automatically selected based on first draft AES. In contrast to many AES systems that achieve high scoring reliability but do not address construct validity (Condon 2013; Perelman 2012), eRevise uses a rubric-based AES system to ensure that dimensions of the construct are well represented by the indicators used for scoring (Loukina et al. 2015). This in turn enables the development of an AWE algorithm for converting internal AES data structures into formative feedback messages that are both tailored to each student's writing needs and aligned to the constructs of the scoring rubric. eRevise is also notable in focusing on evidence usage rather than on surface writing features, and on upper elementary rather than middle or high school students, which makes the application of NLP techniques particularly challenging.

The next two sections describe the eRevise workflow, and the NLP techniques supporting eRevise's AES and AWE components, respectively. This is followed by a classroom deployment and evaluation section demonstrating the promise of eRevise in supporting essay revision. Our deployment tested whether eRevise helped students: 1) improve the overall quality of their drafts when evaluated by human scorers using the RTA evidence rubric, and 2) increase the quantity and relevance/specifity of their text evidence usage when evaluated using NLP. Analyses of 143 essays created by 5th and 6th grade students from 7 different classes support both hypotheses.

System Usage and Architecture

During the first class, a teacher reads an article aloud (with students following along on a hardcopy) about an effort by

[1] Although eRevise currently focuses only on the Evidence rubric, the full RTA as scored by humans comprises five dimensions: Analysis, Evidence, Organization, Academic Style, and MUGS (Mechanics, Usage, Grammar, Spelling).

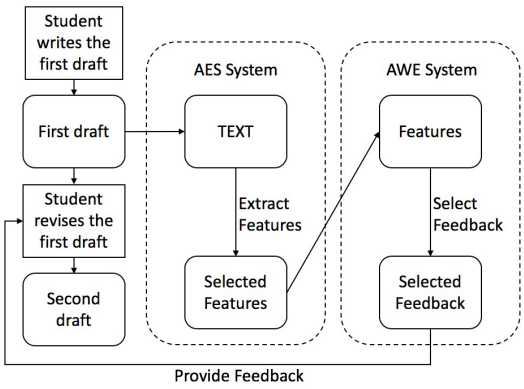

Figure 1: The architecture of the eRevise system.

the Millenium Villages Project (MVP) to eradicate poverty in a Kenyan village.[2] After the teacher discusses predefined vocabulary and asks standardized questions at designated points, there is a prompt at the end of the article which asks students: "Based on the article, did the author provide a convincing argument that winning the fight against poverty is achievable in our lifetime? Explain why or why not with 3-4 examples from the text to support your answer." At this point students use eRevise to write an essay in response to the prompt. Both the source article and the prompt appear on the screen, with students typing their drafts into a text area below the prompt. The purpose of this first usage of eRevise is to electronically collect students' first drafts.

Figure 1 shows the architecture of eRevise. After students submit their first drafts, eRevise's AES component uses a previously developed RTA Evidence scoring algorithm (Zhang and Litman 2017) to extract features representing the quality of text-based evidence usage in terms of constructs in the RTA Evidence rubric. Some of these features (described in the next section) are then passed as input to the AWE system's feedback selection algorithm, which will in turn output a subset of predefined feedback messages that are believed to best address the problems of the first draft based on the features. These formative feedback messages (although not the AES Evidence scores themselves) will be shown to students during the second class period.

During the second class period, students login to eRevise and now revise their first drafts using eRevise. Figure 2 shows a screenshot illustrating revision guided by formative feedback. The left top box shows a student's first draft. This helps students to recall their first drafts and eases revising (e.g., by allowing cutting and pasting). The right-hand side of the screen shows the feedback on the first draft that was automatically selected by the AWE system. The left bottom box shows where students create their second drafts, hopefully guided by the feedback displayed on the right.[3]

[2]While the RTA has three forms (i.e., articles), eRevise currently only supports AES for RTA_{MVP}.

[3]After a student clicks submit, the AES system also scores the revised version of the student's essay. Although eRevise does not share AES scores with students (due to its focus on formative feedback rather than summative assessment), AES scores are included in summary reports later shared with teachers.

Essay Analysis and Feedback Selection

The ultimate goal of our research is to dynamically generate formative feedback that incorporates excerpts from students' essays. However, to simplify the first version of eRevise, the current AWE system instead dynamically selects two of four pre-defined feedback messages to guide students in revising first drafts. Table 1 shows these messages, ordered by a typical progression of evidence usage in student writing development. The messages reflect research on effective feedback and conceptual frameworks for effective text evidence use (Wang, Matsumura, and Correnti 2018), and were created by content experts after an analysis of previously scored RTA essays (Correnti et al. 2013; Rahimi et al. 2014; 2017; Zhang and Litman 2017). These 4 messages were then organized into groups of two based on the appropriateness of the messages for essays with differing evidence sophistication (messages 1 and 2 for the least sophisticated, messages 2 and 3 for more sophisticated, and messages 3 and 4 for the most sophisticated essays). Based on AES feature analysis of evidence usage in the first draft (and further feature processing, described below), each student thus receives two feedback messages based on the group assigned to the essay.

AES Feature Extraction

We have developed several AES systems for RTA assessment (Rahimi et al. 2017; Zhang and Litman 2017; 2018). Our first model (denoted by $Rubric$) (Rahimi et al. 2017) used NLP to represent an essay in terms of features that largely correspond to cells in the RTA Evidence rubric. This rubric, as well as the correspondence between the rubric and features that serve as input to the scoring model, are shown in Table 2. A subsequent model (denoted by SG) (Zhang and Litman 2017) introduced skip-gram word embeddings into the feature extraction process, in order to increase robustness by moving from lexical to semantic similarity computation. Most recently, Zhang and Litman (2018) developed a neural network model with a co-attention layer (denoted by *CO-ATTN*) to eliminate human feature engineering. Table 3 shows performance figures for each of these AES models when evaluated using cross-validation on a previously collected RTA_{MVP} corpus of 2970 essays. Although the neural *CO-ATTN* model has the best performance, to select formative feedback messages that address essay weaknesses in terms of rubric constructs, a more interpretable representation of the essay is necessary. Therefore, SG is the AES system used in eRevise. In particular, two of the features used by SG for score prediction, namely Number of Pieces of Evidence (NPE) and Specificity (SPC), form the basis of eRevise's feedback selection algorithm.[4] Table 4 shows an example first and second draft (with AES Evidence scores of 2 and 3, respectively), along with NPE and SPC values.

Number of Pieces of Evidence (NPE) is an integer encoding the number of evidence topics in the article that are

[4]Although Concentration (CON) is also aligned with the rubric, the other two features are more aligned with the feedback and they are more consequential for improving evidence usage.

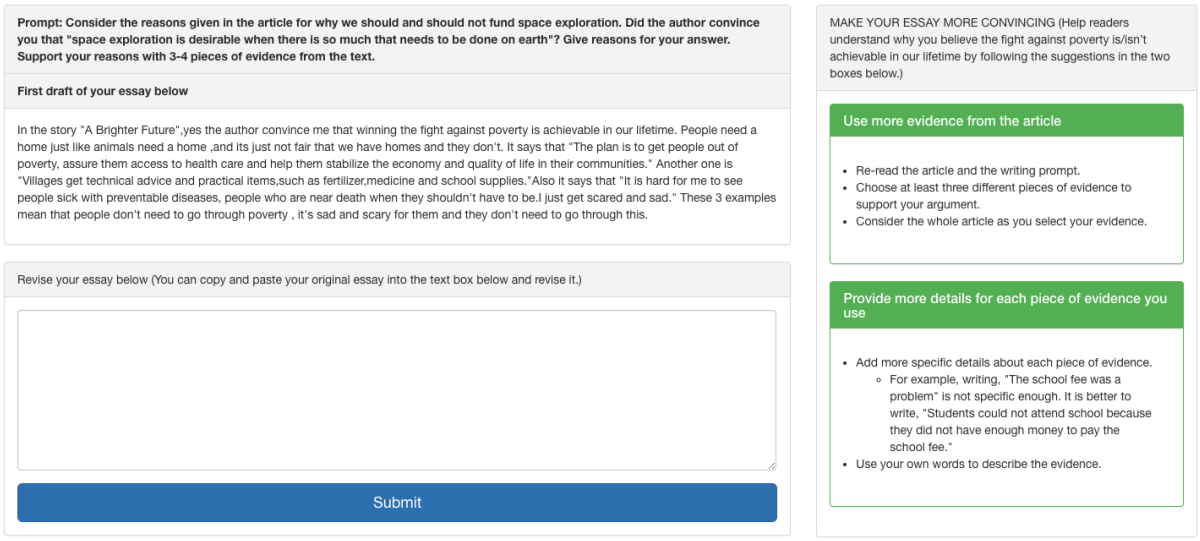

Figure 2: A formative feedback screenshot of the eRevise system.

No.	Name	Feedback
1	Use more evidence from the article	•Re-read the article and the writing prompt. •Choose at least three different pieces of evidence to support your argument. •Consider the whole article as you select your evidence.
2	Provide more details for each piece of evidence you use	•Add more specific details about each piece of evidence. –For example, writing, "The school fee was a problem" is not specific enough. It is better to write, "Students could not attend school because they did not have enough money to pay the school fee." •Use your own words to describe the evidence.
3	Explain the evidence	•Tell your reader why you included each piece of evidence. Explain how the evidence helps to make your point.
4	Explain how the evidence connects to the main idea & elaborate	•Tie the evidence not only to the point you are making within a paragraph, but to your overall argument. •Elaborate. Give a detailed and clear explanation of how the evidence supports your argument.

Table 1: Four feedback messages predefined by content experts, based on progression of evidence use.

mentioned in the essay. A fixed size sliding window algorithm is used to extract this feature. If a window[5] contains at least two similar words from a manually crafted list of main topics and associated words from the article[6], the window is determined to contain text-based evidence related to the topic. Word embedding is used to calculate word similarity, with two words considered as similar after thresholding, thus enabling both lexical and semantic matching (e.g., a student's use of "power" will match "electricity" in the article). In Table 4, the NPE features indicate that the student used text evidence from more topics after revision, i.e., AES identifies one topic (Hospital) in the first draft versus three (Hospital, Farming, and Malaria) in the revised draft - although Malaria is actually a false positive.

Specificity (SPC) is a vector of integers that encodes the number of specific article examples mentioned in the essay. The length of this vector is the number of manually crafted categories, which is 8 for the RTA article (Rahimi et al. 2017). A window-based algorithm is again used for feature extraction, now using a different manually crafted list of words associated with examples and categories from the article, where all examples are assigned to different categories. For example, in the sliding window "*see people sick with preventable diseases*", the essay words "*preventable*" and "*diseases*" match the article word list ("*malaria common disease preventable treatable*") for one of the 6 examples associated with SPC category 4[7]. Therefore, the algorithm increments the value of SPC4 by 1.

AWE Feedback Selection

Although the SPC values (which count the number of times the student mentions specific examples from the article) were useful for developing the AES system via supervised

[5]In eRevise, all windows are of size 6, which optimized AES performance on previously scored training essays.

[6]For the RTA article, the 4 topics are Hospital, Malaria, Farming, and School (Rahimi et al. 2017). Computing similarity with pre-defined topics is typical in content-based AES (Liu et al. 2014).

[7]This category talks about malaria before the MVP program.

	1	2	3	4
Number of Pieces of evidence	Features one or no pieces of evidence (NPE)	Features at least 2 pieces of evidence (NPE)	Features at least 3 pieces of evidence (NPE)	Features at least 3 pieces of evidence (NPE)
Relevance of evidence	Selects inappropriate or irrelevant details from the text to support key idea (SPC); references to text feature serious factual errors or omissions	Selects some appropriate and relevant evidence to support key idea, or evidence is provided for some ideas, but not actually the key idea (SPC); evidence may contain a factual error or omission	Selects pieces of evidence from the text that are appropriate and relevant to key idea (SPC)	Selects evidence from the text that clearly and effectively supports key idea
Specificity of evidence	Provides general or cursory evidence from the text (SPC)	Provides general or cursory evidence from the text (SPC)	Provides specific evidence from the text (SPC)	Provides pieces of evidence that are detailed and specific (SPC)
Elaboration of Evidence	Evidence may be listed in a sentence (CON)	Evidence provided may be listed in a sentence, not expanded upon (CON)	Attempts to elaborate upon evidence (CON)	Evidence must be used to support key idea / inference(s)
Plagiarism	Summarize entire text or copies heavily from text (in these cases, the response automatically receives a 1)			

Table 2: Rubric for scoring the Evidence dimension of RTA. The abbreviations in the parentheses identify features used by the AES system that are aligned with specific assessment criteria (Rahimi et al. 2017).

AES Model	QWK
Rubric	0.632
SG	0.653
CO-ATTN	0.697

Table 3: Quadratic Weighted Kappa (QWK) of different AES models. The *CO-ATTN* model significantly outperforms the *Rubric* and *SG* models, respectively ($p < 0.05$).

machine learning, we found them to be less useful for developing a feedback selection algorithm because the count included duplicate cases, and because the use of word-embedding meant false positive examples were identified during AES. The AWE system thus calculates a new feature named SPC_Total_Merged, which is a count of the number of non-duplicate, unique article examples from the SPC feature vector. For example, in the sentence "**for me to see people sick** with preventable diseases", bolding shows the first example found by the algorithm (window-size is 6, matched words are "people" and "sick"), while underlining shows the second (matched words are "preventable" and "diseases"). While the SPC feature considers these as 2 examples, SPC_Total_Merged considers them as 1 unique example. For the first draft in Table 4, the algorithm thus reduces the SPC total of 11 (from AES, equation below) to a smaller merged total of 6 (for AWE).

After extracting the above features for our previously collected corpora of scored essays, AWE feedback selection was guided by three assumptions: 1) the NPE feature indicates the breadth of unique topics, 2) the SPC_Total_Merged feature indicates the number of unique pieces of evidence the student located and used; and 3) a matrix of these two indicators could match each essay to appropriate feedback. Given we did not need to modify NPE, the following equations were used to calculate SPC_{AWE} for feedback selection.

$$SPC_{total} = \sum_{i=1}^{N} SPC_i \tag{1}$$

SPC_{total}, where N is the number of categories in SPC, calculates the total number of matches the computer finds between students' first drafts and examples we are looking for.

$$SPC_{important} = \sum_{i=S}^{E} SPC_i \tag{2}$$

$SPC_{important}$, where S and E are the start and end indices of important categories, calculates the total number of matched examples from four primary topics for evidence usage (hospital, malaria, farming and school). In the RTA_{MVP}, content experts identified these categories as primary because they are the topics on which students can provide specific examples of improvement, while other SPC categories refer to more general examples from the article.

$$DR = \frac{SPC_{total} - SPC\_Total\_Merged}{SPC_{total}} \tag{3}$$

DR calculates the duplication rate of matched examples, by using SPC_Total_Merged to calculate the proportion of duplicate evidence from SPC_{total}.

$$SPC_{AWE} = RND(SPC_{important} * (1 - DR)) \tag{4}$$

SPC_{AWE} adjusts the number of important matched examples by the duplication rate. This produces a new score for generating feedback, representing the number of unique matched examples from four primary topics. We round the number to get an integer used in the conditional statement below.

$$SPC_{lmh} = \begin{cases} L, & \text{if } SPC_{AWE} < 3 \\ M, & \text{if } SPC_{AWE} \geq 3 \text{ and } SPC_{AWE} \leq 5 \\ H, & \text{otherwise} \end{cases} \tag{5}$$

SPC_{lmh} is a categorical variable for SPC_{AWE} that indicates low (L), medium (M), or high (H) values.

Finally, the AWE system uses NPE (computed during AES) and SPC_{lmh} to select the two most appropriate feedback messages for the essay based on Table 5. The content experts used a previously scored corpus (Zhang and Litman 2017) as development data to manually design this table.

For the first draft in Table 4, $NPE = 1$, $SPC_{total} = 11$, $SPC\_Total\_Merged = 6$, $SPC_{important} = 6$, $DR = 0.455$. $SPC_{AWE} = 3$, and $SPC_{lmh} = M$. Therefore, after

First Draft	**Essay**	In the story "A Brighter Future",yes *the author convince me that winning* the fight against poverty is achievable in our lifetime. People need a home just like animals need a *home ,and its just not fair* that we have homes and they don't. It says that "The plan is *to get people out of poverty*, assure them access to health care and help them stabilize *the economy and quality of life* in their communities." Another one is "*Villages get technical advice and practical* items,such as fertilizer,medicine and school supplies."Also it says that "It is hard *for me to see people sick with preventable diseases*, people who are near death when they shouldn't have to be.I just get scared and sad." These 3 examples mean that people don't need to go through poverty , it's sad and scary for them and they don't need to go through this.									
	Features	NPE 1	SPC1 1	SPC2 2	SPC3 1	SPC4 3	SCP5 0	SCP6 1	SCP7 1	SPC8 2	SPC_Total_Merged 6
Second Draft	**Essay**	In the story "A Brighter Future" yes *the author convince me that winning* the fight against poverty is achievable in our lifetime. Yes we need to win the fight against poverty because everybody needs a home, shelter, food,and money. It ***say that "Their crops were dying because they could not afford*** the necessary fertilizer and irrigation"Another one is that "Its hard *for me to see people sick with preventable diseases,people who are near* death when they shouldn't have to be."Also "***.Little kids were wrapped in cloth*** on their ***mothers backs,or running around in bare*** feet and tattered clothing." These three examples mean that we need to ***help them have a better life*** and a better home than the busty,dirty ground.									
	Features	NPE 3	SPC1 2	SPC2 1	SPC3 2	SPC4 3	SCP5 2	SCP6 1	SCP7 0	SPC8 1	SPC_Total_Merged 6

Table 4: Examples of a student's first and second essay drafts, showing the NLP analyses during AES that are needed for AWE. For each essay, the text-based evidence identified during AES that is used to compute the essay's SPC values is shown in italics (and in bold if only identifed in the second draft). eRevise would display feedback messages 1 and 2 for the first essay draft.

Feature	**Value**														
NPE	0	0	0	1	2	3	4	1	1	2	2	3	4	3	4
SPC_{lmh}	L	M	H	L	L	L	L	M	H	H	M	M	M	H	H
Feedback Messages	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	2,3	2,3	2,3	3,4	3,4

Table 5: Lookup table for feedback selection.

consulting Table 5, eRevise would display feedback messages 1 and 2 for this essay (as for the essay displayed in Figure 2).

In sum, the AWE process results in all students receiving two (of four possible) feedback messages that are selected based on the AES feature analysis and are thus targeted to improving the quality of each student's particular essay. Note that students will receive feedback even when AES predicts a score of 4 for the first draft. In most cases, such students will receive the third and fourth feedback messages focusing on evidence elaboration.

Experimental Deployment and Results

Our first deployment of eRevise took place in two public rural parishes in Louisiana. Seven 5th and 6th-grade teachers had all students in one of their English Language Arts classes write and revise an essay using eRevise. A total of 143 students completed all tasks. We test two hypotheses:

H1: eRevise helps students improve the overall quality of their drafts, as evaluated by human scorers using the RTA evidence rubric.

H2: eRevise increases the quantity and relevance/specificity[8] of evidence that students use from the RTA source text, as evaluated using NLP features.

[8]Corresponding to row 1 and rows 2-3 of Table 2.

The outcome measure for testing H1 is a human-produced RTA Evidence score. After the deployment, a trained human grader used the rubric from Table 2 to score all essays, without knowing whether an essay was a first or second draft. A paired t-test comparing the first and second draft Evidence scores (n=143) supports H1, as the scores improved from first ($MEAN = 2.62$, $SD = 0.95$) to second ($MEAN = 2.72$, $SD = 0.92$) drafts with trending statistical significance ($p \leq 0.08$). The grader also scored essays for the other four RTA dimensions (recall footnote 1). In contrast to Evidence (for which eRevise provided formative feedback to guide revision), there were no significant or trending score improvements for any of these other RTA dimensions (all $p \geq .29$). Finally, the scatter plot in Figure 3a shows that the overall improvement in the Evidence dimension was observed despite potential ceiling effects: 28 students received the maximum score of 4 on their first drafts, 16 of whom also received the maximum on their second drafts. The plot also shows that although the scores increased for 34 students, the scores did not change for the majority of students (and less often even decreased).

We thus explore the use of more fine-grained outcome measures that have a stronger relationship to the eRevise feedback that guided student revision. To test H2, we use the NPE and SPC_Total_Merged features as automatically computed by eRevise during its deployment to ap-

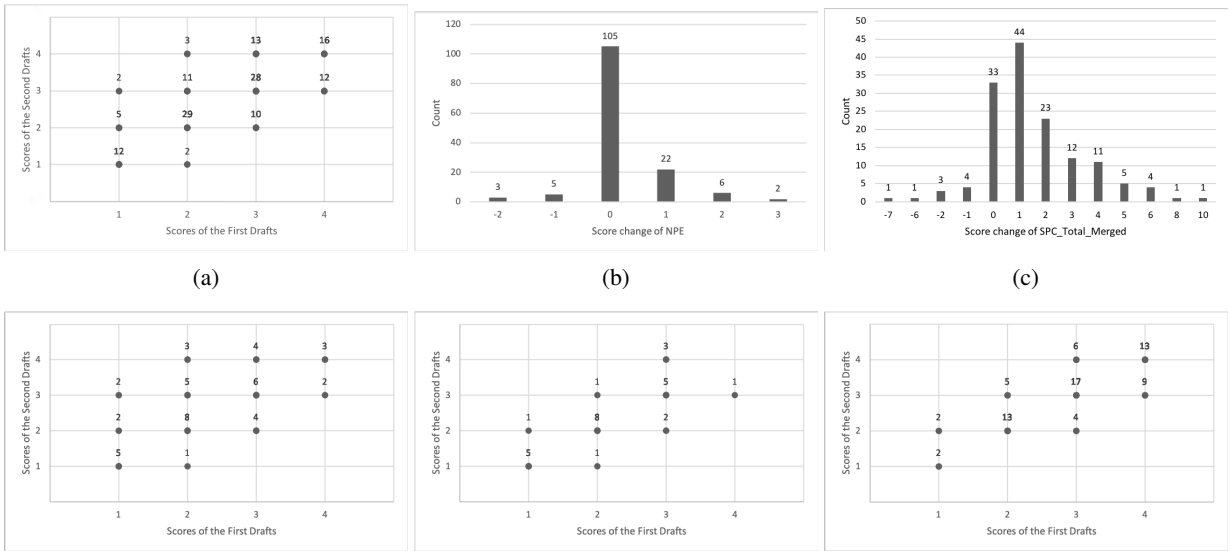

Figure 3: (a) RTA Evidence scores before and after revision. (b) Value changes for the NPE feature. (c) Value changes for the SPC_Total_Merged feature. (d) RTA Evidence scores for essays receiving feedback messages 1 and 2. (e) RTA Evidence scores for essays receiving feedback messages 2 and 3. (f) RTA Evidence scores for essays receiving feedback messages 3 and 4.

proximate evidence quantity and relevance/specificity, respectively. Paired t-tests (n=143) for both support H2. The NPE feature values improved significantly ($p \leq 0.003$) from first ($MEAN = 2.61$, $SD = 1.27$) to second draft ($MEAN = 2.81$, $SD = 1.08$). The SPC_Total_Merged feature values also improved significantly ($p \leq 0.001$) from first ($MEAN = 9.65$, $SD = 4.94$) to second drafts ($MEAN = 11.15$, $SD = 5.39$). For NPE, the histogram in Figure 3b shows that more students added rather than removed evidence (30 versus 8 students). Although 105 students showed no evidence change, 43 were already at ceiling with NPE values of 4 in the first draft. For SPC, the histogram in Figure 3c shows that a large majority of students (101) increased the number of specific article examples that they incorporated into their essays. 33 other students showed no change, while only 9 students removed specific examples.

Recall the 16 students in Figure 3a who were at ceiling when the RTA Evidence score was used as the outcome measure. By instead using the SPC_Total_Merged values as the outcome, these 16 students can now be seen to show improvement from their first drafts ($MEAN = 12.69$, $SD = 4.63$) to the second drafts ($MEAN = 13.25$, $SD = 5.20$), with trending statistical significance ($p \leq 0.095$).

Finally, Figure 3d shows how evidence scores changed for the 45 essays receiving feedback messages 1 and 2. The evidence score improvements from first ($MEAN = 2.33$, $SD = 0.93$) to second ($MEAN = 2.64$, $SD = 0.98$) drafts were statistically significant ($p = 0.02$). Figure 3e shows the score changes for the 27 essays receiving feedback messages 2 and 3. The evidence scores only slightly improved from first ($MEAN = 2.22$, $SD = 0.85$) to second ($MEAN = 2.26$, $SD = 0.94$) drafts. Figure 3f shows that for the 71 essays receiving messages 3 and 4, the evidence scores were almost the same from first ($MEAN = 2.94$, $SD = 0.89$) to second ($MEAN = 2.94$, $SD = 0.81$) drafts. These three figures suggest that drafts with the least sophisticated evidence usage had the most room for improvement. It is also interesting to relate these three feedback-based groupings to essay RTA Evidence scores. 40.63% of drafts receiving Evidence scores of 1 or 2 received feedback messages 1 and 2. 62.03% of drafts receiving Evidence scores or 3 or 4 received messages 3 and 4. Although only 27 essays received messages 2 and 3, 71.43% of these drafts received Evidence scores of 2 or 3.

Current and Future Directions

We are about to begin the next deployment of eRevise, which will extend our work in two ways. First, to better determine the benefit of using AES to adaptively guide revision, we have added a control condition where eRevise will display the same generic feedback message to all students: "MAKE YOUR ESSAY MORE CONVINCING - Help readers understand why you believe the fight against poverty is/isn't achievable in our lifetime." This is in contrast to the existing eRevise adaptive feedback, where students receive different messages based on AES. Second, students will use eRevise for two different forms of the RTA (i.e., RTA_{space} in addition to RTA_{MVP}). While we have already trained a *SG* model for scoring RTA_{space} (Zhang and Litman 2017), Table 5 needs to be verified to ensure the validity of our feedback selection algorithm. We are also exploring adding *CO-ATTN* scores to the lookup table.

For the longer term, we plan to extend our research in other ways. To score a new RTA form, human effort is currently necessary to define topical components, e.g., creating a list of topics and a list of examples for scoring RTA_{space}.

While we have developed pilot data-driven methods that can extract such topical components automatically (Rahimi and Litman 2016), our methods need to be improved so that they do not degrade *SG* model performance. eRevise will also be enhanced to provide feedback for Organization, a second substantive RTA writing dimension for which we already have a pilot AES (Rahimi et al. 2017). We also plan to move from feedback selection to more personalized feedback generation, and to create a teacher dashboard which can automatically generate summaries such as Figure 3a. Finally, since eRevise's feedback encourages students to add more concrete examples from the article, some students may simply copy and paste examples rather than use their own words as the feedback suggests. While the human RTA rubric (last row in Table 2) addresses plagiarism, eRevise currently does not. We thus plan to incorporate the detection of different types of adversarial essays into AES.

Conclusions

eRevise is an AWE system for text evidence usage that uses NLP features produced by a rubric-based AES system to automatically select formative feedback messages most appropriate to a student's needs. By increasing access to feedback on a substantive and important writing dimension, eRevise has the potential to reduce demands on teachers and to build students' knowledge of effective text evidence usage. We first described how eRevise uses NLP techniques to evaluate draft essays and to select appropriate formative feedback messages to guide later revision. Experimental results from a first deployment in 7 classrooms showed that eRevise helped students improve their text evidence usage after receiving formative feedback and engaging in essay revision.

Acknowledgments

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160245 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Chapelle, C. A.; Cotos, E.; and Lee, J. 2015. Validity arguments for diagnostic assessment using automated writing evaluation. *Language testing* 32(3):385–405.

Condon, W. 2013. Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing* 18(1):100–108.

Correnti, R.; Matsumura, L.; Hamilton, L.; and Wang, E. 2013. Assessing students' skills at writing analytically in response to texts. *Elementary School Journal* 114:142–177.

Dikli, S. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment* 5(1):4–35.

Foltz, P. W., and Rosenstein, M. 2015. Analysis of a large-scale formative writing assessment system with automated feedback. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 339–342. ACM.

Graham, S.; Harris, K. R.; and Santangelo, T. 2015. based writing practices and the common core: Meta-analysis and meta-synthesis. *The Elementary School Journal* 115(4):498–522.

Liu, L.; Brew, C.; Blackmore, J.; Gerard, L.; Madhok, J.; and Linn, M. 2014. Automated scoring of constructed-response science items prospects and obstacles. *Educational Measurement: Issues and Practice* 33(2):19–28.

Loukina, A.; Zechner, K.; Chen, L.; and Heilman, M. 2015. Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 12–19.

Perelman, L. 2012. Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring. *International advances in writing research: Cultures, places, measures* 121–131.

Rahimi, Z., and Litman, D. 2016. Automatically extracting topical components for a response-to-text writing assessment. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 277–282.

Rahimi, Z.; Litman, D. J.; Correnti, R.; Matsumura, L. C.; Wang, E.; and Kisa, Z. 2014. Automatic scoring of an analytical response-to-text assessment. In *International Conference on Intelligent Tutoring Systems*, 601–610. Springer.

Rahimi, Z.; Litman, D.; Correnti, R.; Wang, E.; and Matsumura, L. C. 2017. Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education* 27(4):694–728.

Roscoe, R. D.; Allen, L. K.; Weston, J. L.; Crossley, S. A.; and McNamara, D. S. 2014. The writing pal intelligent tutoring system: Usability testing and development. *Computers and Composition* 34:39–59.

Shermis, M. D., and Burstein, J. C. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Shermis, M. D., and Burstein, J. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Wang, E.; Matsumura, L. C.; and Correnti, R. 2018. Student writing accepted as high-quality responses to analytic text-based writing tasks. *The Elementary School Journal* 118(3):357–383.

Weigle, S. C. 2013. English language learners and automated scoring of essays: Critical considerations. *Assessing Writing* 18(1):85–99.

Zhang, H., and Litman, D. 2017. Word embedding for response-to-text assessment of evidence. In *Proceedings of ACL 2017, Student Research Workshop*, 75–81.

Zhang, H., and Litman, D. 2018. Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 399–409.