An Automated Writing Evaluation System for Supporting Self-monitored Revising

Diane Litman(✉), Tazin Afrin, Omid Kashefi, Christopher Olshefski, Amanda Godley, and Rebecca Hwa

University of Pittsburgh, Pittsburgh, PA 15260, USA
{dlitman,taa74,kashefi,ca048,agodley,hwa}@pitt.edu

Abstract. This paper presents the design and evaluation of an automated writing evaluation system that integrates natural language processing (NLP) and user interface design to support students in an important writing skill, namely, self-monitored revising. Results from a classroom deployment suggest that NLP can accurately analyze where and what kind of revisions students make across paper drafts, that students engage in self-monitored revising, and that the interfaces for visualizing the NLP results are perceived by students to be useful.

Keywords: Writing · Revision · Natural language processing

1 Motivation

Automated writing evaluation (AWE) systems driven by natural language processing (NLP) are designed to provide formative feedback for students to revise, and ideally improve, their essays. However, although students do attempt to revise their essays in response to AWE feedback, student revisions often do not yield substantive essay improvements [12,15]. We envision that an enhanced automated writing evaluation system that analyzes and provides feedback on students' revision attempts can support the development of this critical skill. This paper presents the design and classroom evaluation of ArgRewrite, an AWE system that integrates NLP and user interface design to support students in *self-monitored revising*. The NLP backend automatically extracts all revised sentences between two paper drafts, then classifies whether the purpose of each revision was to make a surface (meaning-preserving) versus content (meaning-altering) change. The frontend uses visual interface components to convey the backend's revision analysis. A classroom deployment suggests that NLP provides accurate feedback and that students meaningfully revise.

Compared to prior research, while some AWE systems may detect revisions, they tend to provide feedback on a single essay draft [7,11,13] rather than on

Supported by the National Science Foundation under Grant #173572.

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13355, pp. 581–587, 2022.
https://doi.org/10.1007/978-3-031-11644-5_52

revisions between drafts. *Revision as a skill* is the target of feedback in our work. and more generally has been identified as an area for AWE research [3,13,14].

Most AWE systems provide feedback on task and performance rather than on process [13,14,16]. While recent work has analyzed keystroke logs [5], we analyze process at the level of sentences. Also, while most AWE systems provide actionable feedback messages [17], ArgRewrite instead visualizes NLP results to help students *self-monitor* their revisions, motivated by research on strategy instruction [4], self-regulation [10], and self-monitoring [16].

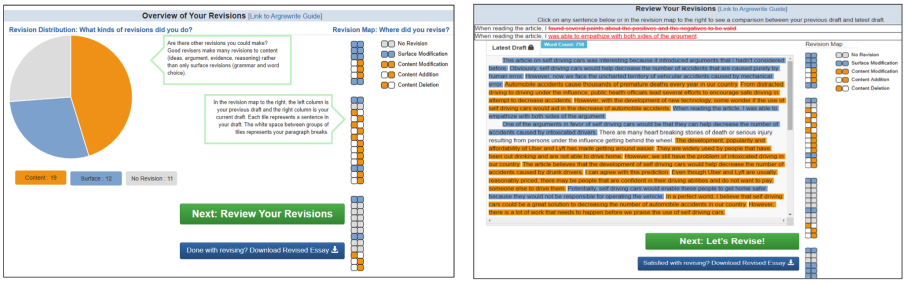
NLP-based writing revision analysis has focused on classifying a revision's purpose [9,19], assessing its quality [2], or understanding temporal patterns [14]. Our *revision extractor and purpose classifier* integrates binary purpose schemas, sentence alignment algorithms, and predictive NLP features from this literature [19]. An alternative approach not requiring revision extraction computes linguistic properties for essay drafts separately, then identifies changes [13].

Prior versions of ArgRewrite were either fully automated demonstration systems not evaluated with users [18] or Wizard of Oz semi-automated prototypes evaluated in lab contexts [1]. Generally, automated NLP revision analysis has not been used to trigger feedback in a target AWE system [13,14]. The ArgRewrite version described below is *fully-automated* and *deployed in a college class.*

2 ArgRewrite: System and Classroom Deployment

The ArgRewrite **backend** uses NLP algorithms developed in our prior work [19] to perform revision extraction and purpose classification. Given the raw text of two essay drafts, an algorithm aligns sentences across drafts based on similarity and global context; remaining added or deleted sentences are aligned to null. The pairs of non-identical aligned sentences are extracted as the essay *revisions.* Finally, a classifier predicts whether the purpose of each revision is to change the meaning of the essay or not [6], using the labels *content* or *surface*, respectively. The classifier was learned using a random forest algorithm due to the small size of the training corpus; four linguistic feature groups encoded each revision [19].

The **frontend** ArgRewrite interfaces were previously evaluated with positive outcomes in a wizarded lab study [1]. The present interfaces were slightly redesigned to better guide students in autonomously using ArgRewrite over the web in a class deployment. Students were first taken to the *Overview Interface* (Fig. 1a), which visualized the NLP results using a revision distribution pie chart (left) and a revision map (right). Next, students were taken to the *Review Interface* (Fig. 1b), which used color coding to convey the purpose labeling for each revised sentence (left). When a student clicked on a sentence revision, a details window showed the character-level differences between the original and revised sentence (top). By clicking the next button, students were taken to the *Revision Interface* (similar to the *Review Interface* but with an additional essay tab with no highlights) to further revise their essay. They were then returned to



(a) Overview Interface (b) Review Interface

Fig. 1. ArgRewrite interfaces

Table 1. ArgRewrite users providing IRB consent (column 1) and their revising summary (column 2). Revisions by students who created draft3 and beyond (column 3). NLP revision purpose classifier performance (column 4).

	Users who upload 2 drafts	User subset who revise (# additional drafts)	Total revisions (avg. per student)	Classifier (F1)
Assignment 1	7	4 (2.3)	26 (8.6)	91.1%
Assignment 2	16	7 (2.7)	291 (41.6)	94.7%
Assignment 3	14	5 (1)	38 (7.6)	90.4%

the *Overview Interface* to start a new cycle of revision (with the re-revised draft automatically uploaded as the latest draft), or to download their final essay.

We **deployed** ArgRewrite in a fall 2019 undergraduate cognitive psychology class that required three writing assignments involving two paper drafts. For each assignment, students 1) wrote draft1 of a paper in response to a prompt, 2) used a peer-review system to provide rubric-guided feedback on the papers of three other students, 3) wrote draft2 of their own paper after receiving peer feedback, and 4) engaged in a final round of peer review. Students were given the option to use ArgRewrite between steps 3 and 4, by submitting draft1 and draft2 of their papers to ArgRewrite, potentially creating further drafts based on the system's feedback, then downloading the final revised draft from ArgRewrite and using it (rather than draft2) for the second phase of peer review.

Although the use of ArgRewrite was completely voluntary, the instructor encouraged it in different ways across assignments: providing a demo in class for Assignment 1, then emailing low-performing students and offering extra credit for Assignment 2. Of the 157 students in the class, 31 used the system at least once. However, only 24 of these students gave IRB consent to use their data for the evaluation below. Of these, 2 students used the system to revise 1 assignment, 19 revised two assignments and only 3 used the system to revise all three assignments. Columns 1 and 2 of Table 1 show the user distribution per assignment.

3 Evaluation and Analysis

Table 2. Revision distributions, using NLP predictions and post-hoc manual annotations. Arrows compare the distributions of AWE versus peer feedback.[a]

	NLP				Manual			
	Draft 1 to 2 (peer feedback)		Draft 2 to final (AWE feedback)		Draft 1 to 2 (Peer feedback)		Draft 2 to final (AWE feedback)	
	Surface	Content	Surface	Content	Surface	Content	Surface	Content
A1	53 (32%)	113 (68%)	14 (54%↑)	12 (46%↓)	43 (27%)	119 (73%)	14 (56%↑)	11 (44%↓)
A2	166 (42%)	231 (58%)	89 (31%↓)	202 (70%↑)	148 (38%)	244 (62%)	94 (33%↓)	191 (67%↑)
A3	139 (51%)	134 (49%)	30 (79%↑)	8 (21%↓)	124 (47%)	138 (53%)	29 (76%↑)	9 (24%↓)

[a] Alignment errors in AWE cause the number of revisions to differ slightly.

NLP Revision Purpose Classifier. NLP performance was analyzed by comparing the classifier's purpose predictions to gold-standard labels that were manually annotated after the assignments were submitted. Each annotation was done by one of three experts familiar with the coding scheme (kappa > .7 in the lab study [1]). The last column of Table 1 shows that for all assignments, macro F1 in binary revision purpose prediction was greater than 90%. This was impressive as the classifier was developed by training on essays responding to two topics from high school English class assignments, but tested on essays responding to three topics from college psychology class assignments.

Student Revision Behavior. Table 1 shows that some students using ArgRwrite engaged in further revision (column 2) beyond peer review (column 1), and often engaged in multiple cycles of revision (column 2 in the parenthesis), e.g., writing 4th and even 5th drafts. Table 1 (column 3) also shows the total number of revisions made by the students who performed self-monitored revising after draft 2. In Table 2, the arrows show that for Assignment 2, the AWE system prompted more content revisions than peer feedback and a high percent of content revisions as compared to surface revisions. Content revisions are generally considered more important in revising and more difficult for students [6]. Possible reasons for the lower percentage of ArgRewrite content revisions in Assignments 1 and 3 could be the students' acclimation to the system during Assignment 1, the direct recruitment with extra credit for Assignment 2, and the low number of students for Assignments 1 and 3. Due to the high accuracy of the NLP classifier, the same inferences (represented by the arrows) can be drawn whether revisions purposes are predicted by NLP or are manually annotated.

Perceived Usability of ArgRewrite. 14 students who used the system for at least assignments 2 and 3 completed a survey at the end of the course. The survey (shown in Table 3) included educational technology usability items (1–7) [8] and items customized for ArgRewrite (8–14). The 'Class' column shows that students responded positively to 13/14 items (i.e., mean Likert values > 3, on a scale from 1–5). The highest score (item 13) indicates that students' perception of classifier

Table 3. Mean scores (1 = strongly disagree; 5 = strongly agree), comparing class deployment (n=14) to lab study (n=22). ** p < .01; * p <.05, $^{+}$ p < .1

	Survey Item	Class	Lab
1	System allows me to have a better understanding of my revision efforts	3.29	3.95*
2	I find the system easy to use	3.43	4.18*
3	My interaction with the system is clear and understandable	3.29	4.14**
4	The system helps me to recognize the weakness of my essay	2.71	3.32
5	System encourages me to make more revisions (quantity)	3.36	3.86
6	System encourages me to make more meaningful revisions (quality)	3.29	3.86
7	Overall the system is helpful to my writing	3.29	3.73
8	I found the "Overview of Your Revisions" page to be useful	3.43	4.14*
9	I found it useful to highlight my revision purposes in different colors	3.71	4.27^{+}
10	I found the revision map visualization useful	3.21	4.09*
11	I found the small window of revision details to be useful	3.43	4.64**
12	I found it helpful to know whether my revision was a "surface" or "content" level change	3.57	4.05
13	My revisions were usually labeled correctly by the system	3.86	4.00
14	I trust the feedback that the system gave me	3.64	3.59

performance reflected the objective results in Table 1. The fact that items 9 and 12 had higher scores than item 10 suggests that feedback on revision purposes was more useful than feedback on revision location. The lowest score (item 4) focused on the essay rather than on the revisions. When focusing specifically on the revisions (e.g., items 1, 5, 6) and the writing process (item 7), the item responses were all positive. We also compared the students in the current study to 22 participants who responded to the same items in our wizarded lab study [1]. The mean 'Class' versus 'Lab' scores were compared using non-paired t-tests, with the results shown in the last two columns of Table 3. This analysis is quasi-experimental since there was no random assignment of survey respondents to the class versus lab conditions. Table 3 shows that for all but the last survey item, the average score in the classroom study was lower than in lab study. Perhaps the class participants are a more critical audience because their actual assignment grade was at stake. Finally, the relative pattern of response values across survey items demonstrated a moderate positive relationship across the class and lab responses (Pearson correlation R=.46, p < .1).

4 Conclusion and Future Directions

This paper described the ArgRewrite system for supporting self-monitored revising. NLP extracts revised sentences between paper drafts and classifies revision purposes, while visualizations convey the NLP results.

A classroom deployment suggests that NLP accurately analyzes revisions and that students engage in self-monitored revising and find the visualizations useful. Future plans include predicting fine-grained purposes using transformers, assessing a revision's quality and alignment with feedback, incorporating system guidance and tutoring, and evaluating via a controlled experiment rather than an 'in the wild' study.

References

1. Afrin, T., Kashefi, O., Olshefski, C., Litman, D., Hwa, R., Godley, A.: Effective interfaces for student-driven revision sessions for argumentative writing. In: Proceedings CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2021)
2. Afrin, T., Litman, D.: Annotation and classification of sentence-level revision improvement. In: Proceedings 13th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 240–246, New Orleans, Louisiana, June 2018
3. Burstein, J., Riordan, B., McCaffrey, D.: Expanding automated writing evaluation. In: Handbook of Automated Scoring, pp. 329–346. Chapman and Hall/CRC (2020)
4. Crossley, S.A., Allen, L.K., McNamara, D.S.: The writing pal: a writing strategy tutor. In: Adaptive Educational Technologies for Literacy Instruction, pp. 204–224, Routledge (2016)
5. Deane, P., Wilson, J., Zhang, M., Li, C., van Rijn, P., Guo, H., Roth, A., Winchester, E., Richter, T.: The sensitivity of a scenario-based assessment of written argumentation to school differences in curriculum and instruction. Int. J. Artif. Intell. Educ. **31**(1), 57–98 (2021)
6. Faigley, L., Witte, S.: Analyzing revision. Coll. Compos. Commun. **32**(4), 400–414 (1981)
7. Foltz, P.W., Rosenstein, M.: Data mining large-scale formative writing. In: Handbook of Learning Analytics, p. 199 (2017)
8. Holden, H., Rada, R.: Understanding the influence of perceived usability and technology self-efficacy on teachers' technology acceptance. J. Res. Technol. Educ. **43**(4), 343–367 (2011)
9. Kashefi, O., et al.: Argrewrite v. 2: an annotated argumentative revisions corpus. Language Resources and Evaluation, pp. 1–35 (2022)
10. MacArthur, C., Philippakos, Z., Ianetta, M.: Self-regulated strategy instruction in college developmental writing. J. Educ. Psychol. **107**(3), 855 (2015)
11. Mayfield, E., Butler, S.: Districtwide implementations outperform isolated use of automated feedback in high school writing. In: International Conference of the Learning Sciences, vol. 2128, London, UK (2019)
12. Roscoe, R.D., McNamara, D.S.: Writing pal: feasibility of an intelligent writing strategy tutor in the high school classroom. J. Educ. Psychol. **105**(4), 1010–1025 (2013)
13. Roscoe, R.D., Snow, E.L., Allen, L.K., McNamara, D.S.: Automated detection of essay revising patterns: applications for intelligent feedback in a writing tutor. Technol. Instr. Cogn. Learn. **10**(1), 59–79 (2015)
14. Shibani, A.: Constructing automated revision graphs: a novel visualization technique to study student writing. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12164, pp. 285–290. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52240-7_52
15. Wang, E.L., et al.: erevis(ing): students' revision of text evidence use in an automated writing evaluation system. Assessing Writing **44**, 100449 (2020)
16. Wilson, J., Huang, Y., Palermo, C., Beard, G., MacArthur, C.A.: Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of mi write. Int. J. Artif. Intell. Educ., 1–43 (2021)
17. Wingate, U.: The impact of formative feedback on the development of academic writing. Assess. Eval. High. Educ. **35**(5), 519–533 (2010)

18. Zhang, F., Hwa, R., Litman, D., B. Hashemi, H.: Argrewrite: a web-based revision assistant for argumentative writings. In: Proceedings f NAACL Conference: Demonstrations, San Diego, California, pp. 37–41 (2016)
19. Zhang, F., Litman, D.: Annotation and classification of argumentative writing revisions. In: Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 133–143. Denver, Colorado, June 2015