International Society of the Learning Sciences

Design of Real-time Scaffolding of Middle School Science Writing Using Automated Techniques

Purushartha Singh, Pennsylvania State University, pxs288@psu.edu
Dana Gnesdilow, The University of Wisconsin-Madison, gnesdilow@wisc.edu
Xuesong Cang, The University of Wisconsin-Madison, xcang@wisc.edu
Samantha Baker, The University of Wisconsin-Madison, srbaker2@wisc.edu
William Goss, The University of Wisconsin-Madison, wgoss2@wisc.edu
ChanMin Kim, Pennsylvania State University, cmk604@psu.edu
Rebecca J. Passonneau, Pennsylvania State University, rjp49@psu.edu
Sadhana Puntambekar, The University of Wisconsin-Madison, puntambekar@education.wisc.edu

Abstract: Science writing skills depend on a student's ability to co-ordinate conceptual understanding of science with the ability to articulate ideas independently, and to distinguish between gradations of importance in ideas. Real-time scaffolding of student writing during and immediately after the writing process could ease the cognitive burden of learning to co-ordinate these skills and enhance student learning of science. This paper presents a design process for automated support of real-time scaffolding of middle school students' science explanations. We describe our adaptation of an existing tool for automatic content assessment to align more closely with a rubric, and our reliance on data mining of historical examples of middle school science writing. On a reserved test set of semi-synthetic examples of science explanations, the modified tool demonstrated high correlation with the manual rubric. We conclude the tool can support a wide range of design options for customized student feedback in real time.

Introduction

Constructing evidence-based scientific explanations is an integral part of science learning (Berland & Reiser, 2009; Krajcik et al., 2014), as emphasized in the Next Generation Science Standards (NGSS; NGSS Lead States, 2013). However, students often write incomplete, non-causal accounts of scientific phenomena (Seah, 2016), and find it difficult to use data appropriately (Sandoval & Millwood, 2005). With technological advances in Natural Language Processing (NLP), recent research has integrated automated scoring of and adaptive feedback on scientific explanations (Gerard et al., 2019; Liu et al., 2016; Riordan et al., 2020; Tansomboon et al., 2017), and on essays in other genres (Zhang et al., 2019), and across languages (Toma et al., 2021). Gerard, Kidron, and Linn (2019) used c-raterML to score 6th graders' scientific explanations in their web-based inquiry science environment and provide feedback for their explanation revisions. They designed automated feedback as well as associated guidance from the teacher to encourage students to use evidence in describing relations among scientific ideas. This work exemplifies *real-time* feedback for students and teachers. Previous studies have developed tools to assess students' scientific writing based on the inclusion of concepts or terms (Liu et al., 2016). Summarization Integrated Development Environment (SIDE) combines NLP and machine learning algorithms to score students' explanations based on the identification of scientific concepts in the responses (Ha et al., 2011). Similar tools include SPSS Text Analysis (SPSSTA) which identifies terms and patterns within students' writing. While these latter systems achieve high score correlations, and sometimes aim for general feedback, they do not provide real-time feedback on specific content elements within science explanation.

In the present study, we aim to create opportunities for students to develop scientific explanations while working on a roller coaster design challenge. We aim to provide real-time feedback that scaffolds students' written explanations of science *concepts* and *relationships* (National Research Council, 2012, p. 52).

An electronic notebook for classroom use

Building on previous implementations of our middle school physics curriculum using paper notebooks (Martin et al., 2019), we have been developing a digital notebook to support students' learning. A digital notebook has multiple advantages over a traditional paper version, such as the ability for students to submit their responses for immediate feedback, use of text mining to provide students with ongoing support, and allowing students to easily revisit previous work. A key feature of the notebook that we are developing is that it can pre-process students' text. Because the digital notebook is accessible through a web browser, students will be able to utilize common features such as spell-checking and grammar-checking, which will make their writing more readable to the teacher, and ensure clean input to an NLP system to support real-time scaffolding. The digital notebook will allow

students an effortless way to reference data, which in turn makes it trivial to remove data from their responses or mark the sentences as being too incomplete for automated feedback.

Through the notebook, students will be given simple feedback based on a combination of data mining and NLP techniques. In our first iteration, data mining will support feedback to students to encourage uniform usage of equations, and to track frequency of key words in their writing. An adaptation of existing content-assessment software (PyrEval) will support feedback on students' written use of science ideas.

Adaptation of PyrEval for real-time scaffolding

PyrEval is a freely available software tool with source code originally designed to assess the content of paragraph-length summaries of source texts, e.g., for reading comprehension (Gao et al., 2019). Characteristics of PyrEval that are a good fit for real-time scaffolding of students' writing of short passages to a prompt are:

1. PyrEval is designed to assess each idea in a student passage for the importance of the idea.
2. It is a lightweight NLP tool that can be used "off-the-shelf" (no domain-specific training is required).
3. It produces **quantitative** scores about importance of ideas, and **qualitative analyses** of specific ideas.

The original PyrEval creates a model of content importance from reference passages written by four to five experts, or by more advanced students. Importance emerges from the frequency of the idea across the references. Emergence, however, does not fit with our objective here, which aim for students to learn specific ideas defined in the curriculum. We modified PyrEval in two ways to address this objective. First, we developed a method to curate a predefined content model based on the curriculum. Second, we adapted the automated analysis to account for features of middle school writing, and to improve the accuracy of the qualitative results.

One significant challenge we faced was to adapt the pre-existing software in the absence of reference texts or a set of examples of the range of inputs the software should handle. In the following subsection, we explain how we mined an existing dataset of passages from middle school students who had a different but related curriculum to synthesize examples like those we expect to see in our new curriculum. We also describe an essay rubric we developed for the new curriculum. The rubric was used both to guide the curation of a new type of PyrEval content model, and to serve as a benchmark against which to measure PyrEval results. Later we discuss the dilemma that arises when we shift PyrEval from reliance on an emergent "*extensional*" definition of content and content importance to also incorporate an "*intensional*" definition constituted by a rubric.

A new rubric with interrater reliability on a semi-synthesized dataset

For this study, we revised historical middle school students' responses to curate data which aligns with our new roller coaster curriculum, by selecting or rephrasing observed statements. We developed a rubric to assess students' writing in the curated essays in line with our expectations for the students. Each time an essay incorporated an explanation of a relevant science concept or relationship between concepts in their essay, it received one point (see examples in Table 1). A total score for each essay was calculated by adding up the total points received. A higher total score indicated that an essay included many important ideas, whereas a score of zero indicated an absence of relevant ideas. Two researchers independently coded a random sample of 13% of the entire sample of 76 essays and achieved a Cohen's Kappa of 0.94, which is considered almost perfect agreement (Stemler, 2001). One of these researchers then coded the remainder of the data.

Table 1

Example of types of concepts and relationships that received one point based on the rubric.

Example Concepts	Example Relationships
• "**Potential energy** is the amount of stored energy an object has based on it's position."	• "the **higher** the initial drop, the **more KE** (kinetic energy) you will have."
• "...the **Law of Conservation of Energy**...states that energy can't be created or destroyed in a system."	• "The more **mass** you have, the **more potential energy** you will have."

Challenges for modifications to PyrEval

PyrEval automatically assesses content by uniquely matching ideas in a student passage to a pyramid model of weighted content units (CUs), and normalizing the summed weights of the matched CUs. PyrEval achieves this through a three-stage pipeline, where the input is a set of reference passages, and student passages (Gao et al., 2019). A pre-processing module decomposes complex sentences into clauses using a rule-based parser and converts the output clauses into semantic vectors. In the second stage, EDUA, a restricted set partition algorithm, groups similar vectors from the reference passages into CUs. In this work, we use EDUA in a new way, to provide data for manual curation of a content model of pre-defined CUs that match a rubric. The third part of the pipeline

applies WMIN, a greedy maximal independent set algorithm (Sakai et al., 2003), to match students' ideas to CUs. For our present work, we investigated which WMIN configurations would perform best.

Table 2

Two distinct configurations of Modified PyrEval from the top 10% of the grid search with the baseline scores the original configuration, and the corresponding correlations. Here, APCS metric refers to the average pairwise cosine similarity, which is the average of the cosine similarity scores of each contributor to the CU with the clause being matched. StDev metric accounts for the standard deviation of the APCS while product metric also accounts for the weight of the CU being matched. These are more coarse grained than just the cosine similarity scores.

PyrEval Configuration	Set B	Set C
(Baseline) SCVP, MinSegLen=1, k=2, s=APCS, m=APCS	0.63	0.78
SCVP, MinSegLen=5, k=2, s=StDev, m=APCS	0.72	**0.85**
FDCP, MinSegLen=3, k=4, s=StDev, m=product	0.70	0.83

Modifications of PyrEval

Three objectives guided our initial experiments to modify PyrEval. The first was to curate a new type of pyramid to achieve a balance between the kind it was originally designed to use and the manual rubric. Our second objective was to investigate how much sentence decomposition would be appropriate for middle school writers. Our past work applied PyrEval to many samples of college writing, achieving high correlations with manual rubrics, but impressionistic comparison of college-level writing with our samples of middle school writing suggested that the latter had more repetition within sentences, and simpler structure. Our third objective was to improve the precision of the automated matching of student statements to pyramid CUs by changing the WMIN scoring parameters. These objectives are highly interdependent, therefore we investigated them in parallel. For example, if a repetitive sentence is decomposed into two segments, and an accurately matching CU is found for one of them, the second segment is likely to get assigned to a CU it does not match as well.

We divided our 76 semi-synthetic essays into three subsets. Subset A (N=10) was used for automatic creation of a variety of pyramid content models that we could mine to populate a manually curated pyramid. Subset B (N=46) was used for tuning PyrEval to middle school science writing, and for refining the curated pyramid. Subset C (N=20) was held in reserve for our final testing of the tuned parameters.

Two features of an emergent pyramid were important to retain. In the original content models, CUs have a Zipfian distribution (Zipf, 1936), which leads to the so-called eighty-twenty rule where twenty percent of distinct item types (here, statements within CUs or essays) account for eighty percent of the observed frequency (i.e., of the same idea stated in different essays), along with a long-tailed distribution of items (i.e., simple clauses) that occur rarely. Pilot experiments indicated that preserving this Zipfian distribution of CUs was important for PyrEval's scoring algorithm (WMIN) to accurately match student statements to the most similar CU, or to no CU. Instead of hard similarity thresholds, WMIN tries to optimize the overall quality of matches. The second important feature of our curated pyramid was to find ***n*** wordings for each important curriculum idea.

By using subset A to automatically generate 10 pyramids, we had multiple data points for defining the distribution of CUs of each weight. We designed the curated pyramid to have the same general distribution. Manual curation of high weight CUs (weights 3-5) was fairly straightforward. We aimed for four weight 5 and four weight 4 CUs, corresponding to the most important relational components of the rubric. To create these, we mined the pyramids that were automatically generated from set A, selecting relevant paraphrases to achieve the desired weight. We found it was also very important to mine the student data for realistic low-weight ideas.

We varied five PyrEval parameters and functions with a different range of values for 675 configurations. We aimed first for high correlation of the scores assigned to the 46 essays in Set B by PyrEval with those assigned by the raters using the rubric. We also inspected random samples of matches between student segments and model CUs to assess the overall quality of the matches as poor, moderate, or good. We identified two configurations representing fairly distinct hypotheses about how to adjust PyrEval for middle school writing, based on three WMIN parameters (top ***k*** matches, node sorting metric (***s***), and match weighting metric (***m***). The minimum number of words in a segment resulting from decomposition (MinSegLen) was also changed to account for the high proportion of short segments in middle school writing.

After choosing two WMIN configurations that performed well on Set B, and that represented different expectations about middle school writing, we tested these configurations on set C. Table 2 shows the two best configurations along with a baseline configuration applied in previous work. Correlations with the rubric are shown for subsets B (development set) and C (test set). The large difference in correlation between sets B and C seems to be due to a few outliers in Set B where a repetitive sentence matched more distinct ideas in the model

that it should have; dropping two outliers raises the Set B correlation to 0.78. Our future work to design custom feedback will investigate both WMIN configurations shown in Table 2.

Conclusion

As modified for a middle school roller coaster curriculum, PyrEval scores based on our curated pyramid model correlate well with manual scores from our highly reliable rubric. In addition, PyrEval provides a qualitative trace of how it computes its scores, which provides us with ratings on each sentence in a student's passage. We have demonstrated that PyrEval can utilize a predefined content model based on the curriculum and have developed a methodology to construct the content models we will need for the curriculum under design.

References

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55. https://doi.org/10.1002/sce.20286

Gao, Y., Chen, S., & Passonneau, Rebecca J. (2019). Automated Pyramid Summarization Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL),* pp. 404-418. Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1038.

Gerard, L., Kidron, A., & Linn, M. C. (2019). Guiding collaborative revision of science explanations. *International Journal of Computer-Supported Collaborative Learning*, *14*(3), 291–324. http://dx.doi.org.ezaccess.libraries.psu.edu/10.1007/s11412-019-09298-y.

Ha, M., Nehm, R. H., Urban-Lurain, M., and Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE-Life Sci. Educ.10* 379-393

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. https://doi.org/10.1002/tea.21299

Martin, N. D., Dornfeld Tissenbaum, C., Gnesdilow, D., & Puntambekar, S. (2019). Fading distributed scaffolds: The interplay between teacher and material scaffolds. Instructional Science, 47(1), 69-98.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, by States*. The National Academies Press.

Riordan, B., Bichler, S., Bradford, A., Chen, J. K., Wiley, K., Gerard, L., Linn, M. C., & Linguist, A. C. (2020). *An empirical investigation of neural methods for content scoring of science explanations*. https://escholarship.org/uc/item/1kf9647q

Sakai, S., Togasaki, M. & Yamazaki, K. (2003). A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Applied Mathematics*, 126(2):313–322. https://doi.org/10.1016/S0166-218X(02)00205-6

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, *23*(1), 23–55. https://doi.org/10.1207/s1532690xci2301_2

Seah, L. H. (2016). Understanding the conceptual and language challenges encountered by grade 4 students when writing scientific explanations. Research in Science Education, 46(3), 413-437. Stemler, S. (2001). An overview of content analysis. Practical assessment, research & evaluation, 7(17), 137-146.

Stemler, S. (2001). An overview of content analysis. Practical Assessment, Research & Evaluation, 7(17), 137-146.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0

Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L.C., Howe E., Qintana, R. (2019). eRevise: Using Natural Language Processing to Provide Formative Feedback on Text Evidence Usage in Student Writing. Proceedings of the 33rd Conference on Artificial Intelligence (AAAI), pp 9619-9625.

Zipf, G. (1936). The Psychobiology of Language. London: Routledge.

Acknowledgements

This work was supported by NSF collaborative awards DRK-12 2010351 and 2010483.