How Well Can You Articulate that Idea? Insights from Automated Formative Assessment

Mahsa Sheikhi Karizaki[1], Dana Gnesdilow[2], Sadhana Puntambekar[2], and Rebecca J. Passonneau[1][0000−0001−8626−811X]

[1] Pennsylvania State University, State College, PA 16801, USA
(mfs6614, rjp49)@psu.edu
[2] University of Wisconsin-Madison, Madison, WI, USA
gnesdilow@wisc.edu, puntambekar@education.wisc.edu

Abstract. Automated methods are becoming increasingly used to support formative feedback on students' science explanation writing. Most of this work addresses students' responses to short answer questions. We investigate automated feedback on students' science explanation essays, which discuss multiple ideas. Feedback is based on a rubric that identifies the main ideas students are prompted to include in explanatory essays about the physics of energy and mass. We have found that students revisions generally improve their essays. Here, we focus on two factors that affect the accuracy of the automated feedback. First, learned representations of the six main ideas in the rubric differ with respect to their distinctiveness from each other, and therefore the ability of automated methods to identify them in student essays. Second, sometimes a student's statement lacks sufficient clarity for the automated tool to associate it more strongly with one of the main ideas above all others.

Keywords: Automated Essay Feedback · Student Writing Clarity

1 Introduction

Science writing has been found to enhance students' inquiry and reasoning skills [5, 8]. Artificial Intelligence has been used to support formative assessment of short answer responses to guide revision [3, 16]. However, using AI tools for revision feedback of science essays is still novel, therefore little is known about accuracy of AI feedback on essays. In a project that provides a web-delivered short curriculum on middle school roller coaster physics, students are prompted to write essays, then revise them based on automated feedback. We find that the feedback accuracy depends on the inherent distinctiveness of propositions that express the main ideas and on how clearly the students express themselves.

The essay feedback comes from PyrEval [2, 13], software that detects main ideas in short passages written to the same prompt. PyrEval can use any pretrained model to convert spans of text to semantic vectors. Before classroom deployment, to optimize accuracy of the feedback, we tested multiple semantic vector methods on a set of manually labelled student data. After classroom use, we manually labeled a new set of essays to assess accuracy on the new classroom sample. By examining patterns of cosine similarities of main idea vectors used as exemplars versus students' main idea vectors, we find that some main ideas are more distinctive than others, and that some student statements have similar cosine similarities to multiple ideas. Both factors that affect PyrEval accuracy.

2 Related Work

Formative feedback, meaning feedback during a unit or course to support further learning, has been found to be most beneficial when it focuses on the *what, how and why* of a problem rather than on verification of results [12]. A series of papers from a group at UC Berkeley have investigated the use of automated guidance in support of short answer explanations from middle school students. They have compared automated feedback alone and in combination with information about the personalized nature of the feedback [14], alone or in combination with students providing feedback on a sample essay [4], and finally alone or in combination with an interface that models the revision process [3]. In all three cases, automated guidance was from the C-rater-ML tool [7], reported to have a 0.72 Pearson correlation with expert humans [14].

Similar investigations by the Concord Consortium [9, 17], mostly with high school students, aimed at improving students' understanding of uncertainty in science [10]. All studies relied on C-rater-ML, achieving QWK scores with humans between 0.78 and 0.93, depending on the study. One study compared generic argumentation feedback to student-specific feedback through use of C-rater-ML, with the latter leading to greater improvements in revisions [17]. Another study compared feedback on argumentation writing alone or in combination with feedback on students' use of science simulations and data [9].

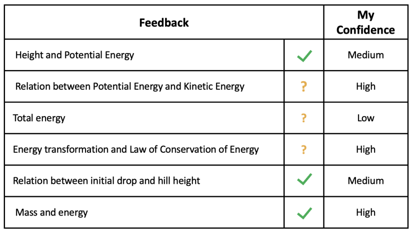
There is relatively little work on automated formative feedback for essay revision. Zhang et al. [16] presented eRevise, which provides rubric-based feedback on students' use of evidence for source-based essays. The authors found that reliance on word embeddings had the best combination of performance accuracy and ability to provide student-specific feedback. Tests with middle school students showed that eRevise led to improved scores on revisions. For our middle school essays, we found that PyrEval performed better using word embeddings rather than contextualized embeddings (cf. section 6), with accuracies from 0.74 to 0.80 across datasets.

3 Roller Coaster Physics Curriculum

During a 2-3 week design-based physics unit, middle school students learned about the physics of energy and energy transfer. They conducted virtual experiments in a web-based environment using a roller coaster simulation, recorded their data, and answered multiple choice and open-ended questions before submitting explanation essays and essay revisions. An essay prompt guided them to explain six ideas, such as the influence of height on potential energy. The six

ID	Sim	Text Description
1	0.69	The greater the height, the greater the potential energy (PE)
2	0.77	As the cart moves downhill, PE decreases and kinetic energy increases
3	0.60	The total energy of the system is always the sum of PE and KE
4	0.75	The law of conservation of energy states that energy cannot be created or destroyed, only transformed
5	0.75	The initial drop should be higher than the hill
6	0.70	Higher mass of the cart corresponds to greater total energy of the system

(a) The Six Main Ideas.



Feedback		My Confidence
Height and Potential Energy	✓	Medium
Relation between Potential Energy and Kinetic Energy	?	High
Total energy	?	Low
Energy transformation and Law of Conservation of Energy	?	High
Relation between initial drop and hill height	✓	Medium
Mass and energy	✓	High

(b) Sample Feedback Checklist: a green check mark means PyrEval detected the idea; a gold question mark indicates PyrEval did not. The 'My Confidence' column reflects PyrEval's average accuracy for a given idea.

Fig. 1: Main Idea content units with average cosine similarities (Sim).

ideas are shown in Fig. 1a, with the type of checklist feedback they might receive in Fig 1b. Elsewhere, we reported that students' revised essays improved based on the automated feedback [11].

4 PyrEval

PyrEval is designed as a lightweight content assessment tool, and is easily adapted to new datasets due to its modular design. Its use of pre-trained semantic vectors means that it requires no training data. Below we explain how we tuned it to our data, including hyperparameter selection. Its three modules are an essay preprocessor, a module to build the content model, and one to assess essays.

The preprocessor converts essays into lists of semantic vectors corresponding to main clauses. This module supports user selection of different semantic vector methods, as we demonstrate in experiments in section 6.

The second module automatically constructs a content model known as a pyramid from five reference essays (exemplars). A pyramid is a list of weighted content units where each content unit represents different ways of expressing the same content extracted from the reference essays. It groups the clause vectors from different exemplar essays into content units (CUs) of at most five vectors. CUs with fewer than five vectors have lower importance. The typical pyramid has a few CUs with the maximum weight of 5, and increasingly more content units for each lower weight. Aligning a pyramid to our rubric is described below.

PyrEval's assessment module [13] constructs a hypergraph graph for each student essay. Each hypernode is a sentence with one internal node per clause, where CUs are associated with internal nodes they are similar to. Edges connect clauses in different sentences that are associated to the same CU. An adaption of a greedy maximal independent set algorithm finds the set of matches (nodes) that give highest sum of CU weights.

Method	*topk*	t	Accuracy
WTMF	3	0.55	0.795
WTMF with MidPhys	3	-0.01	0.675
WTMF Refinement	3	-0.01	0.705
BERT	3	0.85	0.752
Fine Tuned BERT	3	0.83	0.752
BERT + WTMF	3	0.83	0.756

Table 1: Comparison of Six Semantic Vector Methods on GT1

5 Data

Two datasets are used here, one to tune PyrEval to the middle school essays, and one to analyze PyrEval accuracy. In year 2 of the project, we selected 7 high quality student essays to construct 21 pyramids to choose from. We labeled 39 additional essays of varying quality for presence of each main idea, which we refer to as Ground Truth 1 (GT1). Three annotators from the project worked independently, then arrived at a consensus labeling, which was updated several times while testing alternative pyramids. We aimed for a pyramid with exactly six content units of weight 5 (the maximum weight) corresponding one-to-one to the six main ideas in the curriculum. After selecting the pyramid with the best performance on GT1, we manually edited the 5 corresponding reference essays to further improve the pyramid. The *Sim* column of Fig.1a shows the average pairwise cosine similarity of the five vectors within each of the six main idea content units in our final pyramid.

In year 3 of the project, original and revised essays from 60 students were labeled, which we refer to as Ground Truth 2 (GT2). Raters examined the PyrEval feedback, and labeled it as correct or incorrect. Inter-rater reliability was measured on 20% of the essays from two researchers working independently. Substantial agreement of Cohen's Kappa = 0.768 was achieved. Then one of the researchers labeled the remainder of the data.

6 Experiments and Results

Our previous work found WTMF, a matrix factorization vector method, to outperform other word embedding methods [6]. Here we compare six additional vector methods: 1) WTMF with its original corpus; 2) WTMF on the original corpus augmented with MidPhys, a dataset consisting of 11,245 constructed responses from middle school students to 55 physics questions; 3) refinement of the WTMF vector space; 4) BERT contextualized vectors [1] ; 5) BERT fine-tuned on MidPhys; 6) concatenation of vectors 4 and 5. For each method, we performed grid search over two PyrEval hyperparameters: t, the threshold cosine similarity value of a student essay vector to a pyramid content unit to be added to the assessment hypergraph, and *topk*, the number of different student essay vectors that can be associated with the same content unit.

Dataset	PAcc.	NAcc.	Acc.	Rec.	Pre.	F1
GT1	80.64	76.56	79.50	92.77	80.20	86.03
GT2-O	73.73	77.14	74.72	88.67	73.72	80.51
GT2-R	77.00	55.32	74.17	91.98	76.99	83.82
GT2	75.53	70.39	74.44	90.50	75.52	82.34
All	76.78	70.05	75.47	91.09	76.71	83.28

Table 2: PyrEval accuracies, recall, precision and F1.

The third method adapts an approach to refine vectors for opposite sentiment words to have lower cosine similarity [15], that relied on a human ranking of sentiment words. We refined the cosine similarities of a set of key physics terms to be more distant, for word pairs like “potential” and “kinetic,” using tf-idf scores computed on the MidPhys corpus to rank words.

Table 1 compares the six semantic vector methods on the original ground truth dataset GT1. Because method 1 had the highest accuracy, we used this method in our project.

Table 2 reports accuracy on the GT2 dataset. That it is somewhat lower than on GT1 is to be expected, given that GT1 was relatively small in size. Accuracies are broken down into positive accuracy (or sensitivity) and negative accuracy (or specificity). PyrEval’s use of a greedy maximal independent set approach optimizes for the highest sum of matched CU weights, thus inherently favors positive over negative accuracy.

Table 3 shows varied accuracy across the six main ideas. (Accuracy ”bins” in the Fig. 1b checklist are based on GT1 results.) In GT2, the ideas PyrEval identifies most accurately are, in descending order, statements that: define the law of conservation of energy (main idea 4), explain the roller coaster initial drop must be higher than the hill that follows (main idea 5), and that greater mass of the cart results in greater total energy (main idea 6). Main idea three accuracy is modest (71.66%), and accuracy is lower for main ideas one and two.

7 Distinctiveness of Ideas and Student Writing Clarity

PyrEval has higher accuracy on the fourth and fifth main ideas (MIs), which we attribute to greater distinctiveness of lexical items used to express them, such as *transformed, transferred* for MI4 and *initial, drop, hill* for MI5. All the other

Dataset	MI 1	MI 2	MI 3	MI 4	MI 5	MI 6
GT1	76.92	82.05	69.23	89.74	71.79	84.62
GT2 O	63.33	56.66	66.66	91.66	86.66	83.33
GT2 R	63.33	61.66	76.66	86.66	86.66	70.00
GT2	63.33	59.16	71.66	89.16	86.66	76.66
All	66.66	64.77	71.06	89.30	83.01	78.61

Table 3: Accuracy on Main Ideas 1-6 (as percentages).

ideas mention energy, potential energy and kinetic energy, which are relatively close in vector space. While the term *mass* is unique to MI6, its embedding is close to energy terms. The distinctiveness of the main ideas can be quantified by average cosine similarities of all pairs of vectors from the main idea content units, as shown in Table 4. Averaging across pairs gives the lowest similarities (greatest distinctiveness) for MI4 (0.27) and MI5 (0.28), moderate for MI3 (0.37), and around 0.43 for MIs 6, 2 and 1. See below for the *Count* column.

Fig. 2 illustrate cases of clauses with $\geq t$ similarity to multiple MIs. The top of the figure shows two phrases that are more poorly written, and that are candidate matches to three main ideas. The lower half of the figure shows two well articulated statements, with $\geq t$ similarity to exactly one main idea. Column 2 of Table 4 shows how often each MI in a given pair is a candidate match for multiple clauses in a student essay. Pairs of main ideas are shown in ascending order of the number of clauses that have a similarity to both MIs above the threshold t. Main ideas 1 and 5 are the most "confusable" for PyrEval, with 1,152 clauses having similarity $\geq t$ to both.

We plotted distributions of cosine similarities of student vectors to main ideas in a random selection of 117 GT2 essays (out of 159), then verified the consistency of our observations on the remaining 42. We selected one plot to show here. We binned essays by number of PyrEval errors into High, (N=58; 0-1 errors), Mid (N=45; 2 errors), and Low (N=14; ≥ 3 errors) accuracy. Clauses from the High and Mid bins had similarities above t for 1.63 main ideas (sd=0.84). Clauses from the Low bin exceeded t for 1.73 main ideas (sd=0.78). When a clause is a candidate match for up to 3 (*topk*) content units, the algorithm is more likely to err. Fig. 3 plots the cosine similarity (x-axis) by number of clause-main idea pairs at that cosine similarity (y-axis) in an accurately assessed essay of average length versus an inaccurately assessed long essay. The accurate essay (darker bars) has a lower count of clauses overall, but more importantly, very few that have a cosine similarity of 0.70 and above. In contrast, the inaccurate essay has about ten times as many at that cosine similarity and above, which increases the chances that the assessment algorithm would select the wrong node.

8 Conclusion

Through error analysis of a software tool that provides formative feedback on students' science explanation essays, we presented two perspectives on distinctiveness of ideas. First, science explanation statements converted to semantic vectors have different degrees of distinctiveness. Second, students' statements of an idea can be more or less clearly articulated. Both factors affect the accuracy of a software tool we employed to provide formative feedback on students' essays.

9 Acknowledgements

This work was supported by NSF DRK award 2010351.

1. an object has the more PE it will have at the top of the drop and the more total energy (low clarity)
2. the stored energy will turn into kinetic energy because of the gravity (low clarity)
3. but, since the law of Conservation of Energy states that energy can not be created or destroyed, the PE, does not just disappear (high clarity)
4. because, based on the data that was collected, the hill height has to be smaller than the initial drop height (high clarity)

ID	1	2	3	4	5	6
Low Clarity Examples						
1	0.63	-	0.52	0.51	-	0.57
2	-	0.58	-	0.53	-	0.52
High Clarity Examples						
3	-	-	-	0.71	-	-
4	-	-	-	-	-	0.69

Fig. 2: Clauses with low versus high clarity, and main ideas they are similar to.

Pair	Count	Avg. Sim.
4-5	5	0.06
1-**4**	6	0.21
3-**4**	16	0.38
3-5	25	0.16
2-**4**	57	0.30
4-6	69	0.40
1-3	158	0.38
5-6	170	0.27
2-3	214	0.39
2-5	472	0.40
1-6	532	0.44
3-6	534	0.54
2-6	802	0.48
1-2	986	0.59
1-5	1,152	0.53

Table 4: Average cosine similarity of all pairs of main ideas.

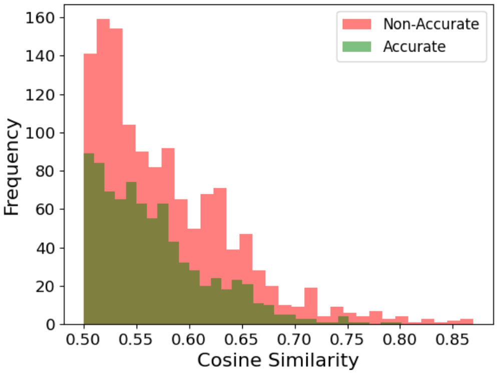


Fig. 3: Cosine similarity distributions of clauses in the full assessment hypergraph for an accurate short essay, and a long inaccurate essay.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 NAACL Conference. pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/N19-1423
2. Gao, Y., Sun, C., Passonneau, R.J.: Automated pyramid summarization evaluation. In: Bansal, M., Villavicencio, A. (eds.) Proceedings of the 23rd CoNLL. pp. 404–418. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/K19-1038

3. Gerard, L., Linn, M.C.: Computer-based guidance to support students' revision of their science explanations. Computers & Education **176**, 104351 (2022). https://doi.org/10.1016/j.compedu.2021.104351
4. Gerard, L., Linn, M.C., Madhok, J.: Examining the impacts of annotation and automated guidance on essay revision and science learnin. In: Looi, C.K., Polman, J.L., Cress, U., Reimann, P. (eds.) Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) (2016)
5. Graham, S., Kiuhara, S.A., MacKay, M.: The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. Review of Educational Research **90**(2), 179–226 (2020). https://doi.org/10.3102/0034654320914744
6. Guo, W., Diab, M.: Modeling sentences in the latent space. In: Proceedings of the 50th Annual Meeting of the ACL. pp. 864–872. Association for Computational Linguistics (2012), https://aclanthology.org/P12-1091
7. Heilman, M., Madnani, N.: ETS: Domain adaptation and stacking for short answer scoring. In: Manandhar, S., Yuret, D. (eds.) Second Joint Conf. on Lexical and Computational Semantics (*SEM), SemEval 2013. pp. 275–279. Assoc. for Computational Linguistics, Atlanta, GA (Jun 2013), https://aclanthology.org/S13-2046
8. Klein, P.D., Boscolo, P.: Trends in research on writing as a learning activity. Jnl. of Writ. Res. **7**(3), 311–350 (2016). https://doi.org/10.17239/jowr-2016.07.03.01
9. Lee, H.S., Gweon, G.H., Lord, T., Paessel, N., Pallant, A., Pryputniewicz, S.: Machine learning-enabled automated feedback: Supporting students' revision of scientific arguments based on data drawn from simulation. Journal of Science Education & Technology pp. 168–192 (2021). https://doi.org/10.1007/s10956-020-09889-7
10. Pallant, A., Lee, H.S., Pryputniewicz, S.: How to support secondary school students' consideration of uncertainty in scientific argument writing. Journal of Geoscience Education **68**(1), 8–19 (2020)
11. Puntambekar, S., Dey, I., Gnesdilow, D., Passonneau, R.J., Kim, C.: Examining the effect of automated assessments and feedback on students' written science explanations. In: Proceedings of the 17th ICLS. pp. 1866–1867 (2023)
12. Shute, V.J.: Focus on formative feedback. Review of Educational Research **78**(1), 153–189 (2008). https://doi.org/10.3102/0034654307313795
13. Singh, P., Passonneau, R.J., Wasih, M., Cang, X., Kim, C., Puntambekar, S.: Automated Support to Scaffold Students' Written Explanations in Science. In: Rodrigo, M., et al. (eds.) Artificial Intelligence in Education, vol. 13355, pp. 660–665 (2022). https://doi.org/10.1007/978-3-031-11644-5“˙64
14. Tansomboon, C., Gerard, L.F., Vitale, J.M., Linn, M.C.: Designing automated guidance to promote productive revision of science explanations. International Journal of Artificial Intelligence in Education **17**, 729–757 (2017). https://doi.org/10.1007/s40593-017-0145-0
15. Yu, L.C., Wang, J., Lai, K.R., Zhang, X.: Refining word embeddings for sentiment analysis. In: Proceedings of the 2017 EMNLP. pp. 534–539. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/D17-1056
16. Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L.C., Howe, E., R., Q.: eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In: IAAI-19 (2019). https://doi.org/10.1609/aaai.v33i01.33019619
17. Zhu, M., Liu, O.L., Lee, H.S.: The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. Computers & Education **143**, 103668 (2020). https://doi.org/10.1016/j.compedu.2019.103668