Automated Support to Scaffold Students' Written Explanations in Science

Purushartha Singh[1], Rebecca J. Passonneau[1], Mohammad Wasih[1], Xuesong Cang[2], ChanMin Kim[1], and Sadhana Puntambekar[2]

[1] The Pennsylvania State University, State College, PA 16801, USA
{pxs288, rjp49, mvw5820, cmk604}@psu.edu
[2] University of Wisconsin-Madison, Madison, WI, USA
xcang@uwisc.edu, puntambekar@education.wisc.edu

Abstract. In principle, educators can use writing to scaffold students' understanding of increasingly complex science ideas. In practice, formative assessment of students' science writing is very labor intensive. We present PyrEval+CR, an automated tool for formative assessment of middle school students' science essays. It identifies each idea in a student's science essay, and its importance in the curriculum.

Keywords: Science explanation · Natural Language Processing

1 Introduction

Secondary school science teachers face multiple demands in scaffolding students' learning of science ideas and science practices. Written explanation of science ideas is an important science practice, as well as a mechanism to assess students' understanding. However, formative assessment of writing is time consuming for teachers. This paper presents a natural language processing application for formative assessment of middle school students' physics essays on energy. It identifies the ideas they express, and the relative importance of these ideas.

Here we first briefly describe current automated support to scaffold science writing. Then we present PyrEval+CR, which extends PyrEval [3], an efficient tool originally developed to assess the content of summaries of the main ideas of source texts.[3] PyrEval+CR has a lightweight, modular design that can be easily adapted to new assignments or writing characteristics. It identifies propositions (statements) expressed in writing, and provides both quantitative and qualitative outputs to support feedback to students and teachers. Section 4 explains how we treat assessment as an optimization problem to match student propositions to propositions in a computable rubric (CR). To evaluate its performance before testing it in the classroom, we constructed a dataset mined from historical essays written by middle school students who used a similar curriculum. An experiment testing many configurations of PyrEval+CR on this data resulted in many settings that correlate well with a highly reliable manual assessment.

[3] PyrEval+CR is available at https://github.com/psunlpgroup/PyrEvalv2.

2 Automated Tools for Scaffolding Science Writing Skills

Previous work points to the potential for automated feedback on student science writing through identification of specific concepts and rubric components. Good agreement between human and automated output has been found in biology explanations from different institutions [6], on middle school explanations of why sugar dissolves in water [7], and on rubric elements for high school biology essays in Hebrew [1]. Below, we present high agreement of our tool with manual rubric scores for the modified middle school essays mentioned above.

Integration of formative assessment tools during science instruction is less well-studied. Teachers who used automated guidance in the WISE environment to help students revise science explanations found that teachers pursued a variety of guidance strategies [10]. A later case study of one teacher's use of automatically generated guidance found the teacher used multiple strategies, and students who revised made more substantial revisions [4]. We have begun a study to apply PyrEval+CR in nearly three dozen middle school classrooms to explore how teachers and students will utilize feedback in classroom settings.

3 PyrEval+CR Overview

PyrEval derives an assessment standard called a pyramid from several reference summaries written by experts. All propositions from the reference summaries are ranked for importance by the number of reference summaries each occurs in. PyrEval+CR relies on a computable rubric with the same form as a pyramid, but derived from a manual rubric. Here we describe the pre-processing that converts an essay to embeddings, and the computable rubric.

The first pre-processing step uses a special-purpose decomposition parser (DP) to decompose complex sentences. DP output consists of alternative ways to decompose the same sentence. For example, a complex sentence of two clauses will have at least two alternatives, one with two clauses, and the original (undecomposed) sentence. Decomposition supports more options for the optimization approach to align student propositions to the CR, as we later illustrate.

The DP uses context-free-grammar parses to extract all tensed verb phrases in a sentence, and dependency parses to identify the subjects of the main verbs. This ensures propositionally complete output clauses. A small set of rules handles traversal of the parses for different syntactic structures [3]. To adapt to middle school writing, we tested subsets of DP rules. We also added a parameter to constrain the minimum length in words of output clauses (MinSegLength).

The second preprocessing step converts DP output clauses to embeddings. Matching a student clause to a CU relies on the average pairwise cosine similarity (APCS) of the student's embedding to sets of CU embeddings. In our earlier work, we found WTMF [5] to give superior similarity results over other embedding methods, but we had not controlled for all factors [3]. Given the widespread use of GloVe [8], we decided to conduct a rigorous comparison between WTMF and GloVe on a standard benchmark, the SemEval semantic textual similarity

Table 1. Pearson correlations with human scores of WTMF and GloVe+SIF on three STS benchmarks. Vocabulary size (V) and total words (S) appear in parentheses.

	WTMF (V=81.8K; S=4M)				Gigaword Sub. (V=67.1K; S=18.9M)			
Test	WTMF		GloVe + SIF		WTMF		GloVe + SIF	
Data	Sent	Win	Sent	Win	Sent	Win	Sent	Win
STS12	**0.7258**	0.6851	0.6859	0.6812	0.6400	**0.6482**	0.6256	0.6256
STS13	**0.7405**	0.6901	0.6426	0.6311	0.5909	**0.6224**	**0.6214**	**0.6214**
STS14	**0.7187**	0.7012	0.6299	0.6149	**0.6835**	**0.6835**	0.6223	0.6223

(STS) tasks (cf. [2]). For STS, humans rated pairs of sentences on a 6-point scale of semantic similarity. System predictions are compared to human ratings using Pearson correlation. We used three years of STS tasks.

WTMF applies weighted matrix factorization to a word-by-sentence matrix of tf.idf scores to compute word embeddings [5]. Using matrix reconstruction, phrase embeddings for unseen sentences can be constructed from the word embeddings. GloVe applies log-bilinear regression to co-occurrence data from a small moving context window over a training corpus [8]. We created GloVe phrase embeddings using a high-performing weighted average of a phrase's word embeddings (SIF) [2]. We trained WTMF and GloVe on a high-quality corpus created by the WTMF developers, and on an extract of the Gigaword news corpus, ensuring both methods used the same vocabulary list for a given corpus. The WTMF corpus combines a high proportion of definitional sentences with a small heterogeneous corpus (the Brown corpus). We sampled increasing amounts of Gigaword, but were unable to achieve matched vocabulary sizes. At nearly five times the size of the WTMF corpus, our Gigaword subset has only 82% of the WTMF corpus vocabulary (see table header in Table 1).

Table 1 shows that WTMF outperforms GloVe+SIF on the benchmark semantic similarity tasks, controlling for the same vocabulary list, corpus, context span, and vector dimensionality (100D). WTMF performs best with the WTMF corpus, using sentence contexts. GloVe results are more consistent across conditions. Due to these results, we use WTMF embeddings from the WTMF corpus.

The original PyrEval creates a set of content units (CUs), called a pyramid, extracted from four to five reference summaries written by experts. Each CU corresponds roughly to a set of paraphrases of the same idea. The number of reference summaries that express the same idea provides an importance weight on the idea. For PyrEval+CR we aimed for an assessment that would more closely resemble the application of an analytic rubric. As described elsewhere [9], we created a very reliable analytic rubric to assess essays that explain students' roller coaster designs with reference to energy concepts (e.g., potential vs. kinetic energy). Here, we describe how we created our computable rubric (CR).

For the CR, we mined phrases corresponding to rubric elements from middle school essays. Figure 1 illustrates a weight 4 CU for a rubric element that defines kinetic energy. In the CR, CU weights range from 5 for important ideas to 1 for weak ideas. The weighted CUs in a PyrEval pyramid have a power law

$\boldsymbol{e}_1$	Kinetic energy is the energy of an object in motion
$\boldsymbol{e}_2$	The energy of an object due to its motion is called kinetic energy
$\boldsymbol{e}_3$	An object that is moving has kinetic energy
$\boldsymbol{e}_4$	Kinetic energy is the energy of a moving object

Fig. 1. A weight 4 (w4) CU. The CR has 62 CUs: 3 w5, 4 w4, 13 w3, 16 w2, 26 w1.

distribution, so we ensure that a CR also does (see Figure 1 caption). All phrases i in a CU are converted to embeddings, represented schematically in column one of Figure 1, as are all decomposed clauses from a student essay.

4 Assessment of Ideas as an Optimization Problem

An independent set for an undirected graph $G = (V, E)$ is defined as a subset $U \subset V$ such that no pair of vertices in U has an edge connecting them in E. A maximal independent set (MIS) is an independent set where no additional vertex from V can be added to U without violating the independent set constraint. The MIS problem is a well-documented NP-complete problem.

PyrEval+CR aims for the optimal way to match student sentences to the CR. Each essay sentence can have several decompositions, and each extracted clause can be more or less similar to each CU. Only one decomposition of a sentence can be used, and each CU can be matched at most once, to penalize repetition. Thus, our assessment task is equivalent to the MIS problem.

WMIN is a greedy weighted MIS algorithm that iteratively adds vertices with the next highest weights to the MIS. At each iteration, all neighbors of a recently added vertex are pruned from the graph. The process repeats until no remaining vertices can be selected. We extended WMIN to operate on a hypergraph (WMIN_H). Each hypernode corresponds to one way to decompose a sentence, and its internal nodes are the extracted clauses and their candidate CU matches. Pruning includes removing other occurrences of a matched CU from the rest of the graph, and recalculating node weights, which use CU weights.

Figure 2 illustrates two sentences in italics, S1 and S2, two of the several decompositions of S1, and the corresponding hypergraph. Hypergraph nodes are labeled by the sentence and decomposition (e.g., S1.2 vs. S1.3), and internal nodes by the clause indices (e.g., S1.2.1, S1.2.2). CU4 (weight 4) from Figure 2 is shown inside the internal node for S1.2.2. Assume that CU5 (weight 2) is a vague statement about kinetic energy and also a potential match for two propositions: clauses S1.3.2 or S.2.1.1. The other internal nodes have anonymous weight 1 CUs (CUX, CUY).

The hypergraph has two edge types. The solid edge between S1.2 and S1.3 constrains selection of at most one decomposition of S1. The dashed edge between internal nodes S1.3.2 and S2.1.1 constrains selection of only one match to CU5. S1.2.2 is a good match to CU4: both state the relation between motion and kinetic energy. The weight of a hypernode is higher if the CU weights are higher, so WMIN_H selects S1.2 over S1.3, S1.3 is pruned, and CU5 only matches S2.1.1.

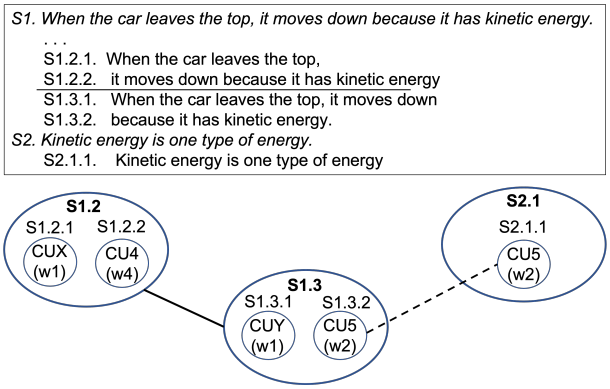

Fig. 2. Illustration of WMIN_H hypergraph nodes and edges.

WMIN_H output for an essay is a log showing the decomposition that was selected for each sentence, and its matched CUs. The essay score is a normalized sum of the weights of the matched CUs.

WMIN_H has parameters to control the greediness of node selection: k for the length of the ranked list of CUs matching each internal node, a sorting metric (s) for ranking this list, and a weighting metric (w) for weighting each hypernode. For both s and w, we tested APCS (see above), the standard deviation of APCS, and the product of APCS and the CU weight (Product).

We tested PyrEval+CR on a curated set of historical essays from a similar curriculum, modified to eliminate sentences that mention ideas not in our current rubric. A set of 76 was subdivided into Set A (N=10) for mining phrases for the CR, Set B (N=46) as a validation set for parameter tuning, and Set C (N=20) for testing. Application of a manual rubric to all 76 is discussed in [9]. We measured performance as the Pearson correlation of the PyrEval+CR score with the manually assigned score. We also reviewed the quality of matches between student's essays and CUs.

We performed grid search on Set B for different subsets of DP rules, different values of MinSegLen (see above), and the three WMIN_H parameters (k, s and w). The DP configurations were all rules (All), all but VP conjunction (-VP) and no decomposition (None). Table 2 reports results for three parameter settings on Sets B and C. DP-VP performed best, but DP-All often worked well. Smaller values of k (k=2, 4) yielded best results, corresponding to a more greedy

Table 2. Example parameter configurations for WMIN_H.

PyrEval Configuration	**Set B**	**Set C**
-VP, MinSegLen=5, k=4, s=StDev, w=APCS	0.69	0.84
All, MinSegLen=3, k=4, s=StDev, w=Product	0.70	0.83
-VP, MinSegLen=5, k=2, s=StDev, w=APCS	0.72	**0.85**

approach that considers fewer CUs per node. Both APCS and Product worked well for w. The standard deviation of APCS usually worked best for s.

The quality of the matches between a randomly selected subset of clauses from set C and CUs was rated by one of the co-authors as poor, moderate, or good. About 93% of the matches were split evenly between moderate and good.

5 Conclusion

PyrEval+CR is intended to support formative assessment for middle school science writing. On a semi-synthetic dataset, the scores correlate very well with a manual rubric. PyrEval+CR produces log output to show which clauses in a student essay match CUs from the computable rubric, along with the relative importance of the CU. Our next steps continue our collaboration with middle school teachers to study how to use PyrEval+CR in a classroom setting.

Acknowledgements NSF DRK12 2010351 and 2010483 funded this work.

References

1. Ariely, M., Nazaretsky, T., Alexandron, G.: Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. Int J Artif Intell Educ **22** (2022). https://doi.org/doi.org/10.1007/s40593-021-00283-x
2. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations (2016)
3. Gao, Y., Sun, C., Passonneau, R.J.: Automated pyramid summarization evaluation. In: Proceedings of the 23rd CoNLL. pp. 404–418 (2019). https://doi.org/10.18653/v1/K19-1038
4. Gerard, L., Kidron, A., Linn, M.C.: Guiding collaborative revision of science explanations. International Journal of Computer-Supported Collaborative Learning **14**(3), 291–32 (2019). https://doi.org/doi.org/10.1007/s11412-019-09298-y
5. Guo, W., Diab, M.: Modeling sentences in the latent space. In: Proceedings of the 50th ACL. pp. 864–872 (Jul 2012), `https://aclanthology.org/P12-1091`
6. Ha, M., Nehm, R.H., Urban-Lurain, M., Merrill, J.E.: Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. CBE—Life Sciences Education **10**(4), 379–393 (2011)
7. Haudek, K.C., Wilson, C.D., Stuhlsatz, M., Donovan, B., Buck-Bracey, Z., Gardner, A., Osborne, J., Cheuk, T.: Using automated analysis to assess middle school students' competence with scientific argumentation. In: National Conference on Measurement in Education. Toronto, ON (2019)
8. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 EMNLP. pp. 1532–1543 (2014). https://doi.org/10.3115/v1/D14-1162
9. Singh, P., Gnesdilow, D., Cang, X., Baker, S., Goss, W., Kim, C., Passonneau, R., Puntambekar, S.: Design of real-time scaffolding of middle school science writing using automated techniques. In: 2022 ISLS (Nov 2022)
10. Tansomboon, C., Gerard, L.F., Vitale, J.M., Linn, M.: Designing automated guidance to promote productive revision of science explanations. Int J Artif Intell Educ **27**(4), 729–757 (2017). https://doi.org/doi.org/10.1007/s40593-017-0145-0