Title

Teacher-informed Expansion of an Idea Detection Model for a Knowledge Integration Assessment

Authors

Li, Weiying
Liao, Yuying
Steimel, Kenneth
et al.

Publication Date
2024-07-09

DOI
10.1145/3657604.3664687

Teacher-informed Expansion of an Idea Detection Model for a Knowledge Integration Assessment

Weiying Li
University of California Berkeley
Berkeley, California, USA
weiyingli@berkeley.edu

Yuying Liao
University of California Berkeley
Berkeley, California, USA
annie.yuying@berkeley.edu

Kenneth Steimel
ETS
Princeton, New Jersey, US
ksteimel@ets.org

Allison Bradford
University of California Berkeley
Berkeley, California, USA
allison_bradford@berkeley.edu

Libby Gerard
University of California Berkeley
Berkeley, California, USA
libbygerard@berkeley.edu

Marcia Linn
University of California Berkeley
Berkeley, California, USA
mclinn@berkeley.edu

ABSTRACT

Students come to science classrooms with ideas informed by their prior instruction and everyday observations. Following constructivist pedagogy, assessments that encourage students to elaborate their ideas, distinguish among them, and link the most promising ones can capture students' potential and help teachers plan their lessons. In this investigation, we study an assessment that engages students in a dialog to refine their response to a Knowledge Integration (KI) question. Our Research Practice Partnership (RPP) initially trained a Natural Language Processing (NLP) idea detection model on 1218 student responses from 5 schools and identified 13 student ideas. The original model had an overall micro-averaged F-score of 0.7634. After classroom testing, three RPP expert teachers with 10+ years of experience reviewed the classroom data and expanded the model, adding six additional ideas including two that they described as precursor ideas because they foreshadowed more sophisticated reasoning. We trained the idea detection model on these 19 ideas using a dataset from 13 teachers and 1206 students across 8 public schools. The updated model had a somewhat lower overall micro-averaged F-score of 0.7297. The two precursor ideas were among the top four detected ideas. The assessment, using the updated model, guided students to express significantly more ideas. A regression model showed that the updated model was associated with greater KI score gains. Expanding the model, thus, created an assessment that motivated students to express more ideas and to achieve higher KI scores. It also provides teachers with deeper insights into their students' understanding of science.

L@S '24, July 18–20, 2024, Atlanta, GA, USA
© 2024 Copyright is held by the owner/author(s).
ACM ISBN 979-8-4007-0633-2/24/07.
https://doi.org/10.1145/3657604.3664687

CCS CONCEPTS

• Applied computing → Education

KEYWORDS

Natural Language Processing, Idea detection, Research Practice Partnership, Adaptive dialog, Rubric refinement, Knowledge Integration

ACM Reference format:
Weiying Li, Yuying Liao, Kenneth Steimel, Allison Bradford, Libby Gerard and Marcia Linn. 2024. Teacher-informed Expansion of an Idea Detection Model for a Knowledge Integration Assessment. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24), July 18–20, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3657604.3664687

1 INTRODUCTION

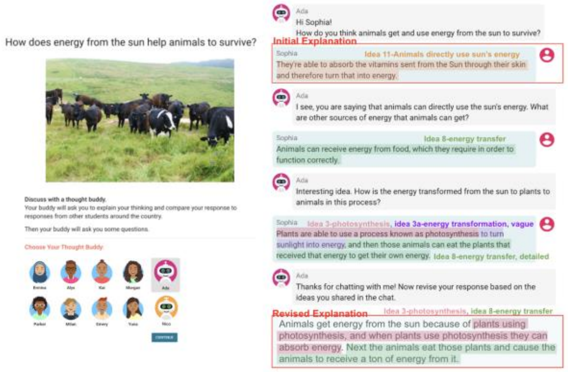
Students' ideas are influenced by their prior instruction and everyday experiences. Constructivist pedagogy emphasizes the importance of assessments that capture and develop these ideas, particularly those grounded in students' prior knowledge, to support teachers in making science accessible [8] and guide students to distinguish among their ideas [5, 6]. Knowledge Integration (KI) assessments encourage students to build on their existing science ideas [6]. In this work, we explore how a web-based dialog with a virtual avatar helps students refine their ideas in a KI assessment. We investigate how a Research Practice Partnership (RPP) consisting of expert middle school science teachers, learning sciences researchers, disciplinary experts, and software designers, leveraged Natural Language Processing (NLP) to review student work, identified new ideas for the NLP model, and refined the KI assessment in web-based dialogs that affirm and build on each student's science ideas. We analyze the value of engaging expert teachers in reviewing the initial model and contributing to a refined NLP idea detection model, thus enhancing the scalability of the assessment [1]. Specifically, we investigate: How can expert teachers expand an idea detection rubric and improve the assessment?

2 RELATED WORK

Language models have been used in formal and informal educational settings for decades to help teachers grade student work and provide tutoring [9]. Recently, NLP has been used to assess students' progress in three-dimensional learning automatically: science and engineering practices, disciplinary core ideas, and cross-cutting concepts [10, 11]. In addition, NLP tools have been implemented in web-based curriculum materials to provide timely adaptive guidance tailored to students' reasoning and synthesize student responses to help teachers monitor class progress [3, 4].

3 THE ENERGY STORY ASSESSMENT

The RPP strengthened an Energy Story Assessment (Figure 1) that asks students to link ideas about energy transfer and transformation during photosynthesis and cellular respiration on an open-sourced Web-based Inquiry Science Environment (WISE). The Assessment includes a dialog featuring adaptive guidance that elicits student ideas and encourages them to improve their response. Students were asked, "How does energy from the Sun help animals to survive?" After responding, students engaged in a dialog with a virtual avatar, described as their thought buddy. We use an NLP idea detection model to automatically identify the ideas in their explanation. We designed adaptive guidance to promote revision of the response, based on one of the ideas detected. After two rounds of adaptive guidance, the student is prompted to use their new ideas to revise their initial explanation.



How does energy from the sun help animals to survive?

Discuss with a thought buddy.
Your buddy will ask you to explain your thinking and compare your response to responses from other students around the country.

Then your buddy will ask you some questions.

Choose Your Thought Buddy

Ada
Hi Sophia!
How do you think animals get and use energy from the sun to survive?

Initial Explanation

Sophia — Idea 11-Animals directly use sun's energy
They're able to absorb the vitamins sent from the Sun through their skin and therefore turn that into energy.

Ada
I see, you are saying that animals can directly use the sun's energy. What are other sources of energy that animals can get?

Sophia — Idea 8-energy transfer
Animals can receive energy from food, which they require in order to function correctly.

Ada
Interesting idea. How is the energy transformed from the sun to plants to animals in this process?

Sophia — Idea 3-photosynthesis, idea 3a-energy transformation, vague
Plants are able to use a process known as photosynthesis to turn sunlight into energy, and then those animals can eat the plants that received that energy to get their own energy. Idea 8-energy transfer, detailed

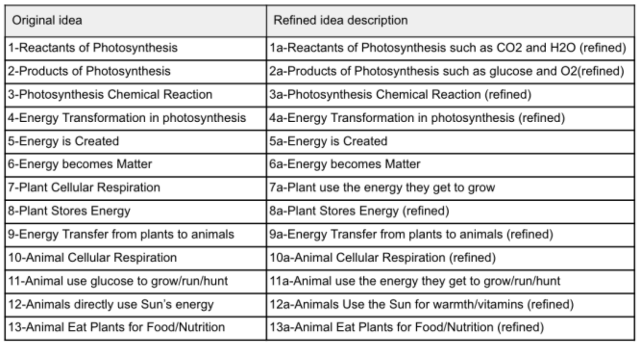
Ada
Thanks for chatting with me! Now revise your response based on the ideas you shared in the chat.

Revised Explanation — Idea 3-photosynthesis, idea 8-energy transfer
Animals get energy from the sun because of plants using photosynthesis, and when plants use photosynthesis they can absorb energy. Next the animals eat those plants and cause the animals to receive a ton of energy from it.

Figure 1: The Energy Story Dialog Assessment. (Each separate idea detected is shown in a different color)

3.1 Initial NLP Idea Detection Model

To develop the initial rubric and label datasets for the NLP idea detection model we collected 1218 student responses to the target question, in prior research, from five schools. Two of the authors (a PhD candidate and a preservice STEM teacher) leading our RPP (learning sciences researchers, middle school science experts, and former biology teachers) identified 13 distinct ideas expressed by students in their responses through inductive coding (Figure 2). Collaboratively, they annotated ideas for 15% of the responses and reached a satisfactory inter-rater reliability (Cohen's Kappa = 0.79). Subsequently, each author individually labeled half of the remaining 85% of responses. The resulting labeled dataset was used to train the initial idea detection NLP model. These 13 ideas include valid science evidence (e.g., energy transfer) and intuitive ideas (e.g., animals eat plants).



Original idea	Refined idea description
1-Reactants of Photosynthesis	1a-Reactants of Photosynthesis such as CO2 and H2O (refined)
2-Products of Photosynthesis	2a-Products of Photosynthesis such as glucose and O2(refined)
3-Photosynthesis Chemical Reaction	3a-Photosynthesis Chemical Reaction (refined)
4-Energy Transformation in photosynthesis	4a-Energy Transformation in photosynthesis (refined)
5-Energy is Created	5a-Energy is Created
6-Energy becomes Matter	6a-Energy becomes Matter
7-Plant Cellular Respiration	7a-Plant use the energy they get to grow
8-Plant Stores Energy	8a-Plant Stores Energy (refined)
9-Energy Transfer from plants to animals	9a-Energy Transfer from plants to animals (refined)
10-Animal Cellular Respiration	10a-Animal Cellular Respiration (refined)
11-Animal use glucose to grow/run/hunt	11a-Animal use the energy they get to grow/run/hunt
12-Animals directly use Sun's energy	12a-Animals Use the Sun for warmth/vitamins (refined)
13-Animal Eat Plants for Food/Nutrition	13a-Animal Eat Plants for Food/Nutrition (refined)

Figure 2: Detected ideas and refinements

We developed and applied an NLP model using a multilabel token classification approach [10]. The NLP model architecture included a pretrained transformer backbone SciBERT [2], which leverages 1.1 million scientific papers for pretraining and vocabulary creation, a bidirectional GRU-based RNN, and a final linear projection, trained using Binary Cross Entropy loss. Hyperparameter tuning, involving epochs and learning rates, was conducted via a 10-fold cross-validation grid search. For KI scores, we applied a previously developed NLP KI scoring model to score the written explanations for KI (scale 1-5) [7], which measures the overall accuracy and coherence of explanations.

The multi-label idea detection model achieved an overall micro-averaged F-score of 0.7634. We designed guidance for each of the 13 ideas that encourages the student to elaborate their reasoning. When multiple ideas were detected, the dialog responded to intuitive ideas first, then infrequent ideas and normative ideas.

3.2 Refining the Model

We asked three expert science teachers who were members of the RPP each with over 10 years of experience teaching WISE units including photosynthesis. They also taught the Photosynthesis unit with the embedded NLP dialog. They reviewed the Energy Story rubric for 20 randomly chosen student dialogs and reflected based on their classroom teaching. They identified 6 additional ideas (Figure 3, idea 14-19) including two precursor ideas "sun helps plants grow" and "sun helps animals grow", two ideas related to prior instruction about food chains "energy conservation" and "energy decrease", and two intuitive ideas "direct negative impact of the sun on individuals" and "animals' direct use of glucose without any transformation."

We investigated the impact of adding the ideas identified by the expert teachers to the NLP model.

4 METHODS

4.1 Participants

The RPP, led by the original rubric designers, refined the rubric using a dataset involving 1206 students taught by 13 teachers across 8 public schools.

4.2 Expanded NLP Idea Detection Model

We randomly selected 400 student dialogs, with the greatest distribution across teachers and the least missing data. These dialogs underwent 4 rounds of annotation, each comprising 100 student dialogs. The first round focused on enhancing distinction between ideas, particularly between Ideas 9 and 13. In the second round, we strengthened the rubric by categorizing the 19 ideas as normative, intuitive, and broad. The third round entailed further refinement of label descriptions (Figure 3). In the final round, two researchers achieved satisfactory inter-rater reliability (Cohen's Kappa = 0.81) in applying the new rubric to label student responses.

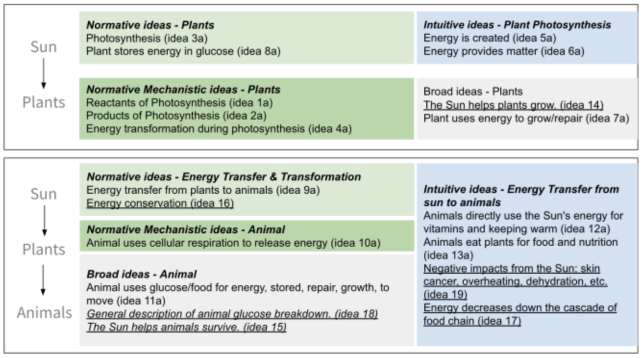

Sun
Plants
Normative ideas - Plants
Photosynthesis (idea 3a)
Plant stores energy in glucose (idea 8a)
Intuitive ideas - Plant Photosynthesis
Energy is created (idea 5a)
Energy provides matter (idea 6a)
Normative Mechanistic ideas - Plants
Reactants of Photosynthesis (idea 1a)
Products of Photosynthesis (idea 2a)
Energy transformation during photosynthesis (idea 4a)
Broad ideas - Plants
The Sun helps plants grow. (idea 14)
Plant uses energy to grow/repair (idea 7a)
Sun
Plants
Animals
Normative ideas - Energy Transfer & Transformation
Energy transfer from plants to animals (idea 9a)
Energy conservation (idea 16)
Normative Mechanistic ideas - Animal
Animal uses cellular respiration to release energy (idea 10a)
Broad ideas - Animal
Animal uses glucose/food for energy, stored, repair, growth, to move (idea 11a)
General description of animal glucose breakdown. (idea 18)
The Sun helps animals survive. (idea 15)
Intuitive ideas - Energy Transfer from sun to animals
Animals directly use the Sun's energy for vitamins and keeping warm (idea 12a)
Animals eat plants for food and nutrition (idea 13a)
Negative impacts from the Sun: skin cancer, overheating, dehydration, etc. (idea 19)
Energy decreases down the cascade of food chain (idea 17)

Figure 3: Updated Idea rubric (NOTE: Underlined ideas are teacher-identified new ideas)

5 RESULTS

5.1 Updated Model Accuracy

After validating the updated rubric and establishing inter-rater reliability, the two researchers individually labeled each distinct idea in half of the remaining 806 student dialogs, forming a labeled dataset to train the updated idea detection NLP model. To assess our model's performance in NLP dialogs, we relied on precision, recall, and the F-score as key metrics. Precision measured the accuracy of positive predictions, ensuring our model identified specific science concepts accurately. Recall evaluated the model's capability to capture all relevant instances, minimizing the risk of missing essential science ideas expressed by students. The F-score, a balanced blend of precision and recall, provided a comprehensive assessment, considering both false positives and false negatives. Using the SciBERT backbone, the updated model achieved an overall micro-averaged F-score of 0.7297 and macro-averaged F-score of 0.4483, indicating a slight decrease in performance compared to the original model's micro-averaged F-score of 0.7634.

Of the 6 new ideas, the two precursor ideas had high accuracy and frequency. Idea 14 (Sun helps plants grow) had an F-score of 0.7370 (precision: 0.75, recall: 0.73, 7547 tokens). Idea 15 (Sun helps animals grow) had an F-score of 0.6922 (precision: 0.69, recall: 0.69, 4619 tokens). The two ideas connected to prior instruction varied in F-scores, precision, and accuracy. Idea 16 (energy conservation) had an F-score of 0.3660 (precision: 0.56, recall: 0.27). Idea 17 (Energy decreases down the food chain) had a F-score of 0.5131 (precision: 0.73, recall: 0.40). Both ideas had higher precision scores than recall scores. The two intuitive ideas had very low F-scores. Idea 19 (The Sun brings skin cancer etc) had an F-score of 0.0565 (precision: 0.29, recall: 0.03). Idea 18 (animals use glucose without any energy transformation) had an F-score of 0.1242 (precision: 0.26, recall: 0.08).

Of the refined 13 ideas, the updated model was better at detecting five ideas (3a, 6a, 8a, 9a, 12a) and about as good for 3 ideas (10a, 11a, 12a) (F-score >0.73). The updated model was less accurate for 4 ideas (1a, 2a, 5a, 7a).

5.2 Impact on KI scores and ideas

We implemented the updated NLP adaptive dialog with 100 6th-grade students taught by one of the expert teachers. We compared their KI scores and idea frequencies with those of 67 7th-grade students using the original NLP dialog taught by the same teacher. Using the dialog with the updated model, students received significantly higher KI scores (M = 2.56, SD=0.29) at their revised explanations after the adaptive dialog compared to their initial explanations (M = 2.22, SD=0.31, t(99) = 5.07, p < .001). Their KI scores improved by 15.3%. Using the dialog with the original model, students also received significantly higher KI scores (M= 2.78, SD = 0.30) at their revised explanations after the adaptive dialog compared to their initial explanations (M = 2.48, SD = 0.34, t(66) = 4.44, p < .001). Their KI scores improved by 12.0%. Students using the updated model improved more than those who used the original model. Using the generalized linear model, we found significant associations between the KI score increase and the predictor variable Model (Original vs Updated) (AIC = 560.3, p < .001). The estimated coefficient for the Updated model compared to the Original model was 0.237 (p < .01), indicating that, holding other variables constant, the updated model was associated with a greater increase in KI score compared to the original model.

For the idea frequency, students expressed more ideas at their revised explanation using the updated dialog ($N_{updated}$=289, $N_{original}$=123, Figure 4) with 6 new ideas accounting for 50.0% of all the expressed ideas. The Poisson generalized linear model revealed significant associations between the count and the predictor variables Model (Original vs Updated) and DialogTime (Initial vs Revised) (AIC = 1031.4, p < .001). The estimated coefficient for the updated model compared to the original model was 0.486 (p < .001), indicating that, holding other variables constant, the updated model was associated with an increase in the expected idea count compared to the original model. In addition, the two precursor ideas were two of the four most detected ideas in the updated model, demonstrating their importance for guiding students.

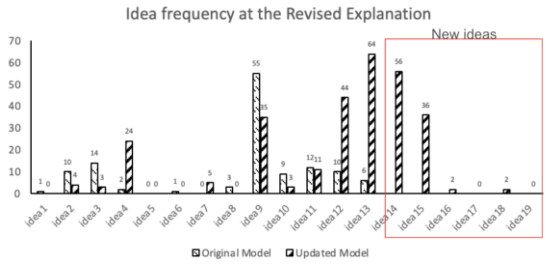


Figure 4: Idea frequency at the revised explanation from students from Teacher A, using the original model compared to using the updated model

6 DISCUSSION

Using an iterative process, our RPP expanded and refined an idea rubric to train an NLP model to engage students in a dialog as part of a KI assessment. Review of the initial model by expert teachers resulted in the addition of six ideas that were not in the original rubric: two precursor, two prior instruction related, and two intuitive ideas. When we included these ideas in the expanded rubric, the updated model illustrated the importance of the two precursor ideas, which were precisely and frequently detected. Further, the updated model had only a slight decrease in micro-averaged F-score from 0.7634 to 0.7297. The dialog enhanced assessment, using the updated model, also enabled students to express more ideas and to achieve higher KI scores, more accurately measuring their potential to integrate their ideas.

The impact of the updated model illustrates the importance of including expert teachers who have classroom experiences with the model in an RPP. It underscores the necessity of continually refining models to align with the diverse ways in which students express their understanding in classrooms.

The initial rubric was built by members of the RPP including middle school science teachers, educational design researchers, disciplinary experts, computer scientists, and software designers, familiar with KI pedagogy and science concepts. The updated rubric benefitted from RPP teachers who had taught the photosynthesis unit and recognized precursor ideas. These ideas, when detected in the assessment, enabled students to elaborate their understanding of the topic.

Although the three expert teachers who taught the photosynthesis unit were able to identify missing ideas when reviewing classroom adaptive dialogs, they may not have anticipated these ideas had they only examined the responses used for developing the initial rubric. The teachers could imagine how they would guide students when they looked at the dialogs generated using the initial rubric. They drew attention to broad precursor ideas that were not identified as distinct ideas by the developers of the initial rubric. These ideas turned out to be especially valuable for guiding students to revise and refine their responses. As we seek to support learning at scale, it is important to regularly revisit the rubrics used for creating NLP models, and the forms of expertise needed to update the rubrics.

Generalization of this work is limited by the specific assessment context, populations of students, and the members of the RPP. It would benefit from replication with different KI assessments, students, and teachers. The value of identifying broad ideas that students express and using a dialog to prompt them to refine these ideas is a promising way to increase the value of open-ended assessments and inform teachers of the reasoning behind students' initial responses. Next steps include exploring ways to anticipate broad ideas in other KI assessments and automate the design of dialogs to elicit more details from students.

REFERENCES

[1] Atteberry, A., Loeb, S. and Wyckoff, J. 2017. Teacher Churning: Reassignment Rates and Implications for Student Achievement. *Educational Evaluation and Policy Analysis.* 39, 1 (Mar. 2017), 3–30. DOI:https://doi.org/10.3102/0162373716659929.

[2] Beltagy, I., Lo, K. and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv.

[3] Dickler, R., Adair, A., Gobert, J., Hussain-Abidi, H., Olsen, J., O'Brien, M. and Pedro, M.S. 2021. Examining the Use of a Teacher Alerting Dashboard During Remote Learning. *Artificial Intelligence in Education* (Cham, 2021), 134–138.

[4] Gerard, L., Matuk, C., McElhaney, K. and Linn, M.C. 2015. Automated, adaptive guidance for K-12 education. *Educational Research Review.* 15, (Jun. 2015), 41–58. DOI:https://doi.org/10.1016/j.edurev.2015.04.001.

[5] Kali, Y. 2006. Collaborative knowledge building using the Design Principles Database. *International Journal of Computer-Supported Collaborative Learning.* 1, 2 (Jun. 2006), 187–201. DOI:https://doi.org/10.1007/s11412-006-8993-x.

[6] Linn, M.C. and Eylon, B.-S. 2011. *Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration.* Routledge.

[7] Liu, O.L., Lee, H.-S., Hofstetter, C. and Linn, M.C. 2008. Assessing Knowledge Integration in Science: Construct, Measures, and Evidence. *Educational Assessment.* 13, 1 (Apr. 2008), 33–55. DOI:https://doi.org/10.1080/10627190801968224.

[8] Luna, M.J. 2018. What Does it Mean to Notice my Students' Ideas in Science Today?: An Investigation of Elementary Teachers' Practice of Noticing their Students' Thinking in Science. *Cognition and Instruction.* 36, 4 (Oct. 2018), 297–329. DOI:https://doi.org/10.1080/07370008.2018.1496919.

[9] Nye, B.D., Graesser, A.C. and Hu, X. 2014. AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education.* 24, 4 (Dec. 2014), 427–469. DOI:https://doi.org/10.1007/s40593-014-0029-5.

[10] Riordan, B., Bichler, S., Bradford, A., King Chen, J., Wiley, K., Gerard, L. and C. Linn, M. 2020. An empirical investigation of neural methods for content scoring of science explanations. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (Seattle, WA, USA → Online, Jul. 2020), 135–144.

[11] Zhai, X., Krajcik, J. and Pellegrino, J.W. 2021. On the Validity of Machine Learning-based Next Generation Science Assessments: A Validity Inferential Network. *Journal of Science Education and Technology.* 30, 2 (Apr. 2021), 298–312. DOI:https://doi.org/10.1007/s10956-020