Identifying NGSS-Aligned Ideas in Student Science Explanations

Brian Riordan[1], **Aoife Cahill**[1], **Jennifer King Chen**[2], **Korah Wiley**[2],
Allison Bradford[2], **Libby Gerard**[2], **Marcia C. Linn**[2]
[1]ETS
[2]University of California-Berkeley

Abstract

With the increasing use of online interactive environments for science and engineering education in grades K-12, there is a growing need for detailed automatic analysis of student explanations of ideas and reasoning. With the widespread adoption of the Next Generation Science Standards (NGSS), an important goal is identifying the alignment of student ideas with NGSS-defined dimensions of proficiency. We develop a set of constructed response formative assessment items that call for students to express and integrate ideas across multiple dimensions of the NGSS and explore the effectiveness of state-of-the-art neural sequence-labeling methods for identifying discourse-level expressions of ideas that align with the NGSS. We discuss challenges for idea detection task in the formative science assessment context.

Students in grades K-12 in the U.S. increasingly engage in learning about science and engineering through online environments that provide learning experiences with interactive simulations and experiments. Teachers use these formative assessments that involve short text responses from students to assess student understanding and thereby fill in knowledge gaps and build on productive ideas. While automated scoring of student responses is a well-studied task (Burrows, Gurevych, and Stein 2015; Pado 2016; Dzikovska, Nielsen, and Leacock 2016; Shermis 2015), effective automated methods to analyze student responses in more detail hold similar potential to reduce the burden on teachers to exhaustively read student responses and allow them to instead focus on targeted student support. For science education, of particular interest is the capability to identify regions of responses that express concepts or display skills that align with standards such as the Next Generation Science Standards (NGSS; NGSS Lead States (2013)).

The NGSS call for the integration of three dimensions of science learning: disciplinary core ideas (DCIs), crosscutting concepts (CCCs), and science and engineering practices (SEPs)[1]. In this work, we describe the design of constructed response items that formatively assess student understanding of multiple NGSS dimensions, namely, using SEPs while demonstrating integrated understanding of DCIs and CCCs. We then explore the effectiveness of state-of-the-art neural sequence-labeling methods for identifying the expression of high-level science and engineering concepts in responses from U.S. middle school students interacting with an online science education environment[2].

Datasets

Students at 11 middle schools in the U.S. engaged in science units in an online classroom and contributed written responses to assessment questions as part of pre- and posttests in the units. We designed three free-response assessment questions embedded in the units that aligned with the NGSS. The questions were designed to elicit student reasoning about two or more NGSS dimensions of ideas and concepts (DCIs and CCCs) and practices (SEPs) (Table 3 in Appendix). Spans of student responses were annotated for ideas related to each of these elicited dimensions.

The questions were from three units (Table 3). The three science units and questions were as follows: (1) The Thermodynamics Challenge (TC) unit asked students to determine the best material for insulating a cold beverage using an online experimentation model. The assessment question asked students for both scientific concepts and to explain proposed experiments. (2) In the unit on photosynthesis and cellular respiration (PH), students interacted with dynamic molecular models and wrote integrated explanations of how photosynthesis supports the survival of both plants and animals. (3) In Solar Radiation (SR), students were asked to agree or disagree with a claim made by a fictional peer about the functioning of a solar oven based on working with an interactive model.

We designed annotation rubrics for each question corresponding to the two question dimensions. The rubrics provided guidance for how the NGSS "performance expectation" for that dimension could be realized by students in the context of answering the question. Specifically, we synthesized the ideas, concepts, and practices described in the NGSS Evidence Statement documents of each targeted performance expectation to develop the annotation criteria. As

[1]https://www.nextgenscience.org/resources/ngss-appendices

[2]https://wise.berkeley.edu/

	#	avg. #	avg. len	avg. % unique wds
TC-Sci	160	0.28	22.2	1.94
TC-Exp	153	0.27	23.2	1.36
PH-CCC	417	0.84	54.3	1.58
PH-DCI	256	0.51	35.8	1.46
SR-Sci	91	0.19	20.8	4.73
SR-Eng	318	0.65	23.6	2.94

Table 1: Descriptive statistics for span annotation, in terms of total number (#), average number per response (avg. #), average number of tokens (avg. len) and average % unique words per span.

	baseline (majority)	baseline (O)	word	word+char
TC-Sci	.104	.297	.519	**.525**
TC-Exp	.105	.297	.557	**.584**
PH-CCC	.256	.178	**.616**	.611
PH-DCI	.098	.214	**.539**	.538
SR-Sci	.073	.310	**.552**	.550
SR-Eng	.219	.219	**.704**	.701

Table 2: Macro-averaged F1 scores for sequence labeling.

an example, Table 4 in the Appendix lists the specific concepts for the TC question.

The statistics in Table 1 give an overview of the characteristics of our data. The size of our datasets is comparable to previous work (SR=492, TC=588, PH=499) (cf. Schulz et al. (2018)). The data is challenging to model in several respects. First, most of the dimensions' data are relatively sparse. The average number of concept spans per response is less than 0.5 for all but one dimension (PH-CCC). Second, all of the data is characterized by long idea spans, with the shortest average span still greater than 20 words. Third, there is significant lexical variability in some concept types (SR-Sci, SR-Eng).

Methods

Task. We formulate the task of identifying spans of student ideas as a token-level sequence labeling task (cf. Schulz et al. (2018)). Spans are labeled with both type and boundaries following the standard BIO scheme. We build independent models for each span type (DCI, CCC, or SEP), aggregating across targeted concepts, resulting in 6 models.

Network architecture. We explored the BiRNN-CRF family of network architectures for this study, which has demonstrated state-of-the-art performance for sequence labeling on similar tasks (Schulz et al. (2018). A bidirectional recurrent network (here, GRU) processes the sequence of tokens. The contextualized representations produced for each token are processed by a Conditional Random Field model over the token labels. We explore the effectiveness of both word- and character+word-based models to partially alleviate noise from spelling variation.

For comparison, we implemented two baselines: (1) predict the *O* tag; (2) predict the most frequent non-*O* tag.

Data preparation and model training. We trained models with 5-fold cross validation with train/dev/test splits. We split the data into 5 folds of 60% train, 20% dev, and 20% test. For hyperparameter tuning, we evaluated performance only on the dev sets and recorded the best performance across epochs. We evaluate performance with macro-averaged F1 score (unweighted) (cf. M_S metric in Schulz et al. (2018)). For training final models after hyperparameter tuning, we combined the training and dev sets and stopped training at the average best epoch across dev folds rounded to the nearest 5th epoch (cf. Johnson and Zhang (2017)). The final test performance was the average test performance across folds. Further details about the data and model are provided in the Appendix.

Results and Discussion

Table 2 displays the models' performance across questions and NGSS dimensions. First, the models outperform the majority class and O-tag baselines. Second, the character+word models perform competitively with – but often don't exceed – the performance of the token-based models, indicating that character representations do not always provide an additive benefit for noisy data on this task. Third, we see substantial variation in F1 scores across NGSS dimensions within the data for each question (e.g., among the word-based models, SR-Sci=.552 while SR-Eng=.702).

As a first step in analyzing the reasons for model performance, we fitted generalized linear mixed-effect models (GLMMs) to the per-response macro-averaged F1 score data with questions as random effects, aggregating across questions and NGSS dimensions. For each response, we computed the response length in tokens, span length, and number of unique tokens in spans (a measure of lexical variability). Surprisingly, we found no significant effect of these predictors. This may indicate that per-response prediction performance may be affected less by high-level statistical properties of the data typically associated with task difficulty. Instead, the interaction of the representations for individual lexical items across the response may drive performance.

We also conducted a manual error analysis of exact span matches. Results suggested that some of the models may have suffered from exposure bias, i.e. often predicting the extremely frequent *O* label. We find that the questions with sparse annotations tended to lead to models with 'missed detections', failing to predict most of the gold-labeled spans. Conversely, in the questions with higher coverage, we find that the models do tend to predict many more non-*O* labels, and as a result many more spans, many of which overlap completely or partially with the gold spans.

In this work, we described the development and annotation of constructed response items for detecting students' ideas aligned with NGSS-defined ideas, concepts, and practices. We found that neural sequence-labeling methods that have proved successful on similar tasks can achieve moderate performance on this task. Future work will explore explicit model features to improve accuracy and methods to explain model predictions to support targeted feedback to students and teachers in formative assessment applications.

ID	Unit	Question dimension 1	Question dimension 2
TC	Thermodynamics Challenge	Science: insulators, conductors and heat energy transfer (DCI)	Experimentation: informative experimental tests and comparisons (SEP)
PH	Photosynthesis	Energy transfer drives matter cycling (CCC)	Photosynthesis and producer/consumer relationships (DCI)
SR	Solar Radiation	Science: Heat energy transfer (CCC)	Engineering (SEP)

Table 3: Datasets and NGSS dimensions.

References

Burrows, S.; Gurevych, I.; and Stein, B. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education* 25(1):60–117.

Dzikovska, M. O.; Nielsen, R. D.; and Leacock, C. 2016. The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications. *Language Resources and Evaluation* 50(1):67–93.

Johnson, R., and Zhang, T. 2017. Deep Pyramid Convolutional Neural Networks for Text Categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *HLT-NAACL*.

Pado, U. 2016. Get semantic with me! the usefulness of different feature types for short-answer grading. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Schulz, C.; Meyer, C. M.; Sailer, M.; Kiesewetter, J.; Bauer, E.; Fischer, F.; Fischer, M. R.; and Gurevych, I. 2018. Challenges in the Automatic Analysis of Students' Diagnostic Reasoning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

Shermis, M. D. 2015. Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment* 20(1):46–65.

States, N. L. 2013. *Next Generation Science Standards: For States, By States*. Washington, D.C.: The National Academies Press.

Zhang, X., and Goldwasser, D. 2019. Sentiment Tagging with Partial Labels using Modular Architectures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Appendix

Data

Here we provide more detail about one of the three datasets. For the Thermodynamics Challenge unit, we designed a constructed response question that aligns with the NGSS performance expectation MS-PS3-3 and assesses student performance proficiency with the targeted DCIs in the performance expectation, understanding of the SEP of planning and carrying out an investigation, and the integration of both of these to construct a coherent and valid explanation. The constructed response question prompts students to explain the rationales behind their experiment plans with the model, using both key conceptual ideas as well as their understanding of experimentation as a scientific practice: "Explain WHY the experiments you [plan to test] are the most important ones for giving you evidence to write your report. Be sure to use your knowledge of insulators, conductors and heat energy transfer to discuss the tests you chose as well as the ones you didn't choose." Table 4 provides the individual concepts that were labeled for each question dimension.

Science: insulators, conductors and heat energy transfer (DCI)
(1) Thermal energy is transferred from hotter objects to colder objects (2) Different materials have different properties AND/OR rates of conductivity (3) Insulators have a lower rate of conductivity/thermal energy transfer AND/OR Conductors have a higher rate conductivity/thermal energy transfer
Experimentation: informative experimental tests and comparisons (SEP)
(1) Need to test insulators (or materials that have a low rate of conductivity or that minimize heat energy transfer (2) Need to run tests in a hot room

Table 4: Thermodynamics Challenge concepts for each question dimension.

Network details

For a succinct overview of neural CRF models with word and character representations for sequence labeling, see Zhang and Goldwasser (2019), Section 3.

The model with additional character representations represents each word with a sequence of 25-dimensional character embeddings (randomly initialized). A character encoder encodes these sequences, and the output for each token is concatenated with the token's word embedding before the word-level encoder.

The data was tokenized with the spaCy tokenizer. For the word tokens, we used GloVe 100 dimension vectors (Pennington, Socher, and Manning 2014) as pretrained embeddings and fine-tuned these during training. Word tokens that were not found in the embeddings were mapped to a randomly initialized UNK embedding.

Networks were trained to maximize the CRF loglikelihood score (Lample et al. 2016). From experiments on our dev sets, the best-performing optimizer was Adadelta with learning rate of 1.0, using a batch size of 32 and gradient clipping set to 1.0. During training, we maintain an exponential moving average of the model's weights. The maximum decay rate is set to 0.999.

Hyperparameter tuning

For the combined word-character encoder, we varied the encoder hidden dimensions in $\{100, 250\}$, number of layers in $\{1, 2\}$, dropout on embeddings in $\{0.0, 0.25\}$. We obtained the best results on average across all datasets with 2 layers, 100 dimensions, and variational dropout of 0.25.

For the character encoder, we used a CNN and varied the number of filters in $\{50, 100\}$ and the filter sizes in $\{3, 5, (3,4,5)\}$ (i.e. the concatenation of filter sizes 3, 4, and 5). For these experiments, we used a combined word-character encoder with the best hyperparameter settings from the word encoder tuning experiments. The best character encoder results were achieved with 100 filters and filter sizes of (3,4,5).