An empirical investigation of neural methods for content scoring of science explanations

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

Abstract

With the widespread adoption of the Next Generation Science Standards (NGSS), science teachers and online learning environments face the challenge of evaluating students' integration of different dimensions of science learning. Recent advances in representation learning in natural language processing have proven effective across many natural language processing tasks, but a rigorous evaluation of the relative merits of these methods for scoring complex constructed response formative assessments has not previously been carried out. We present a detailed empirical investigation of feature-based, recurrent neural network, and pre-trained transformer models on scoring content in real-world formative assessment data. We demonstrate that recent neural methods can rival or exceed the performance of feature-based methods. We also provide evidence that different classes of neural models take advantage of different learning cues, and pre-trained transformer models may be more robust to spurious, dataset-specific learning cues, better reflecting scoring rubrics.

1 Introduction

The Next Generation Science Standards (NGSS) call for the integration of three dimensions of science learning: disciplinary core ideas (DCIs), cross-cutting concepts (CCCs), and science and engineering practices (SEPs) (NGSS Lead States, 2013). Science teachers can promote knowledge integration of these dimensions using constructed response (CR) formative assessments to help their students build on productive ideas, fill in knowledge gaps, and reconcile conflicting ideas. However, the time burden associated with reading and scoring student responses to CR assessment items often leads to delays in evaluating student ideas. Such delays potentially make subsequent instructional interventions less impactful on student learning. Effective automated methods to score student responses to NGSS-aligned CR assessment items hold the potential to allow teachers to provide instruction that addresses students' developing understandings in a more efficient and timely manner and can increase the amount of time teachers have to focus on classroom instruction and provide targeted student support.

In this study, we describe a set of CR formative assessment items that call for students to express and integrate ideas across multiple dimensions of the NGSS. We collected student responses to each item in multiple middle school science classrooms and trained models to automatically score the content of responses with respect to a set of rubrics. This study explores the effectiveness of three classes of models for content scoring of science explanations with complex rubrics: feature-based models, recurrent neural networks, and pre-trained transformer networks. Specifically, we investigate the following questions:

(1) What is the relative effectiveness of automated content scoring models from different model classes on scoring science explanations for both (a) holistic knowledge integration and (b) NGSS dimensions?

(2) Do highly accurate model classes capture similar or different aspects of scoring rubrics?

2 Methods

2.1 Background

We focus on constructed response (CR) items for formative assessments during science units for middle school students accessed via an online classroom system (Gerard and Linn, 2016; Linn et al., 2014). In past research, items that assessed NGSS performance expectations (PEs) were scored with a single knowledge integration (KI) rubric (Liu

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 135–144
July 10, 2020.

et al., 2016). KI involves a process of building on and strengthening science understanding by incorporating new ideas and sorting out alternative perspectives using evidence. The KI rubric used to score student short essays rewards students for linking evidence to claims and for adding multiple evidence-claim links to their explanations (Linn and Eylon, 2011). In this study, we develop items that solicit student reasoning about two or more NGSS dimensions of DCIs, CCCs, and SEPs. We score each item for KI and NGSS "subscores" relating to the DCIs, CCCs, and practices.

3 Scoring item and rubric design

In this section we describe the design of the CR items that comprise the datasets for the content scoring models. The CR items formatively assess student understanding of multiple NGSS dimensions, namely, using SEPs while demonstrating integrated understanding of DCIs and CCCs.

We designed formative assessment items and associated rubrics for four units currently used in the online classroom system: Musical Instruments (MI), Photosynthesis and Cellular Respiration (PS), Solar Ovens (SO), and Thermodynamics Challenge (TC).

Musical Instruments and the Physics of Sound Waves (MI). The Musical Instruments unit engages students in testing and refining their ideas about the properties of sound waves (wavelength, frequency, amplitude, and pitch) and guides them in applying what they learn to design and build their own instrument, a water xylophone. The CR item we designed aligns with the NGSS PE MS-PS4-2 PE and assesses students' understanding of the relationship of pitch and frequency (DCI) and the characteristics of a sound wave when transmitted through different materials (CCC). Students are prompted to distinguish how the pitch of the sound made by tapping a full glass of water compares to the pitch made by tapping an empty glass. In their answer, they are asked to explain why they think the pitch of the sound waves generated by striking the two glasses will be the same or different.

Photosynthesis and Cellular Respiration (PS). This unit engages students in exploring the processes of photosynthesis and cellular respiration by interacting with dynamic models at the molecular level. We designed a CR item that aligns with NGSS performance expectation MS-LS1-6 that asks students to express an integrated explanation of how photosynthesis supports the survival of both plants and animals. This item explicitly solicits students' ideas related to the CCC of matter cycling (i.e. change) and energy flow (i.e. movement): "Write an energy story below to explain your ideas about how animals get and use energy from the sun to survive. Be sure to explain how energy and matter move AND how energy and matter change." Successful responses demonstrate proficiency in the SEP of constructing a scientific argument and reflect the synthesis of the DCIs and CCCs.

Solar Ovens (SO). The Solar Ovens unit asks students to collect evidence to agree or disagree with a claim made by a fictional peer about the functioning of a solar oven. Students work with an interactive model where they explore how different variables such as the size and capacity of a solar oven affect the transformation of energy from the sun. We designed a CR item that addresses NGSS PE MS-PS3-3 and assesses students for both the CCC of energy transfer and transformation and the SEP of analyzing and interpreting data. After working with the interactive model, students respond to the CR item with the prompt: "Explain why David's claim is correct or incorrect using the evidence you collected from the model. Be sure to discuss how the movement of energy causes one solar oven to heat up faster than the other."

Thermodynamics Challenge (TC). The Thermodynamics Challenge unit asks students to determine the best material for insulating a cold beverage using an online experimentation model. We designed a CR item that aligns with the NGSS PE MS-PS3-3 and assesses student performance proficiency with the targeted DCIs in the PE, 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 CR item prompts students to explain the rationale 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."

We designed three scoring rubrics for each item corresponding to two "subscores" representing the degree to which the written responses expressed

PE-specific ideas, concepts, and practices and one KI score that represents how the responses integrated these elements.

NGSS subscore rubrics. To evaluate the written responses for the presence of the DCIs, CCCs, and SEPs, we designed subscore rubrics for two of the three dimensions (Table 1). Specifically, we synthesized the ideas, concepts, and practices described in the "evidence statement" documents of each targeted performance expectation to develop the evaluation criteria. We assigned each response a score on a scale of 1 to 3, corresponding to the absence, partial presence, or complete presence of the ideas, concepts, or practices.

KI score rubrics. The ideas targeted by the KI scoring rubrics aligned with subsets of the ideas described in the evidence statements. For example, the KI scoring rubrics for the Photosynthesis item evaluated written responses for the presence and linkage of five science ideas related to energy and matter transformation during photosynthesis. KI rubrics used a scale of 1 to 5.

3.1 Data collection

Participants were middle school students from 11 schools. Students engaged in the science units and contributed written responses to the CR items as part of pre- and post-tests. Across schools, 44% of students received free or reduced price lunch and 77% were non-white.

All items were scored by two researchers using the item-specific subscore and KI and rubrics described above. To ensure coding consistency, both researchers coded at least 10% of the items individually and resolved any disagreements through discussion. After the inter-rater reliability reached greater than 0.90, all of the remaining items were coded by one researcher (cf. the procedure in Liu et al. (2016))[1].

Table 2 displays the dataset sizes and mean words per response for the KI scores and NGSS subscores, and Figure 1 depicts the respective score distributions. Among the holistic KI scores, the highest score of 5 had relatively fewer responses than other score levels. By examining the shape of the distributions of scores across the NGSS subscores, we can see that students' expression of different aspects of NGSS performance expectations differed across items. For the Musical Instruments and Photosynthesis items, students expressed the disciplinary core ideas less than the cross-cutting concepts. For both the Solar Ovens and Thermodynamics Challenge items, students often did not explicitly articulate science concepts. The Thermodynamics Challenge item was particularly challenging, as many students did not express the targeted science or experimentation concepts.

[1]Datasets are not publicly available because of the IRB-approved consent procedure for participants (minors) in this research.

Item	PE	DCI	CCC	SEP
Musical Instruments	MS-PS4-2	•	•	
Photosynthesis	MS-LS1-6	•	•	
Solar Ovens	MS-PS3-3		•	•
Thermodynamics Challenge	MS-PS3-3	•		•

Table 1: NGSS performance expectations (PE) and targeted components: disciplinary core idea (DCI), cross-cutting concept (CCC), and science and engineering practices (SEP) targeted by each item.

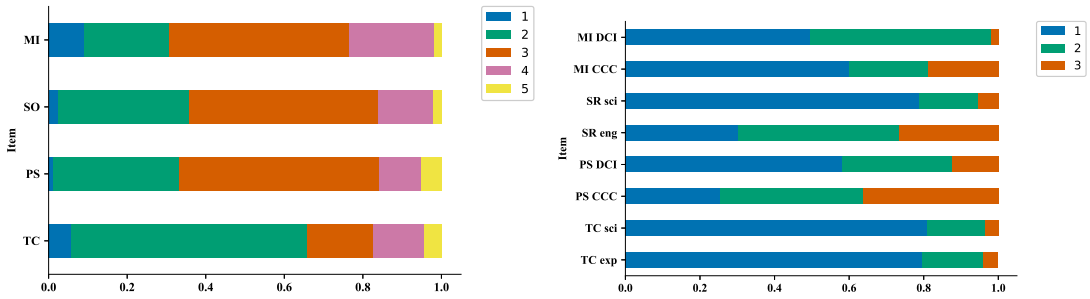
Item	Type	Responses	Mean words per response
MI	KI	1306	25.40
PS	KI	1411	54.57
SO	KI	1740	31.87
TC	KI	994	31.73
MI	CCC DCI	1306	25.40
PS	CCC DCI	553	70.40
SO	SEP: eng CCC: sci	605	32.62
TC	SEP: exp DCI: sci	583	31.43

Table 2: Descriptive statistics for each item's dataset.

3.2 Content scoring models

Content scoring models were built for each item and score type (knowledge integration and two NGSS dimensions). Models for each score type were trained independently on data for each item. In this way, the three models for an item formed different "perspectives" on the content of each response. Human-scored training data for the NGSS dimension models comprised either a subset of or overlapped with the training data for the KI models.

The models were trained to predict an ordinal score from each response's text, without access to expert-authored model responses or data augmentation. This type of "instance-based" model (cf. Horbach and Zesch (2019)) is effective when model responses are not available and can score responses of any length without additional modeling



(a) (b)

Figure 1: Score distributions for (a) knowledge integration scores and (b) NGSS subscores.

complexity. As we focus on content scoring, the models do not consider grammatical or usage errors that do not relate to the content of each response.

The feature-based model is a nonlinear support vector regression (SVR) model. The model is trained on a feature set of binarized word n-grams with n in $\{1, 2\}$.

The RNN model uses a simple architecture with pre-trained word embeddings and pooling of hidden states. Pre-trained word embeddings are processed by a bidirectional GRU encoder. The hidden states of the GRU are aggregated by a max pooling mechanism (Shen et al., 2018). The output of the encoder is aggregated in a fully-connected feedforward layer with sigmoid activation that computes a scalar output for the predicted score. Despite its simplicity, this architecture has achieved state-of-the-art performance on benchmark content scoring datasets (Riordan et al., 2019).

For the pre-trained transformer model, we used a standard instance of the BERT model (Devlin et al., 2019). BERT is a bidirectional transformer model trained on the tasks of masked token prediction and next sentence prediction across very large corpora (BooksCorpus and English Wikipedia). During training, a special token '[CLS]' is added to the beginning of each input sequence. To make predictions, the learned representation for this token is processed by an additional layer with nonlinear activation, outputting a score prediction. The model was 'fine-tuned' by training the additional layer's weights on each item's dataset.

3.3 Data preparation, model training, and hyperparameter optimization

SVR model. The SVR models used an RBF kernel. Hyperparameters *C* and *gamma* were tuned on the validation sets and were optimized by root mean squared error.

RNN model. Word tokens were embedded with GloVe 100 dimension vectors (Pennington et al., 2014) and fine-tuned during training. Word tokens that were not found in the embeddings vocabulary were mapped to a randomly initialized UNK embedding. On conversion to tensors, responses were padded to the same length in a batch; these padding tokens are masked out during model training. Prior to training, responses were scaled to [0, 1] to form the input to the networks. The scaled scores were converted back to their original range for evaluation.

The GRUs were 1 layer with a hidden state of size 250. The RNN models were trained with a mean squared error loss. For this investigation, the RNN was optimized with RMSProp with ρ of 0.9, learning rate 0.001, batch size 32, and gradient clipping (10.0). We used an exponential moving average of the model's weights for training (decay rate = 0.999) (Adhikari et al., 2019). In the tuning phase, models were trained for 50 epochs.

Pretrained transformer model. We used the *bert-base-uncased* pre-trained model (Wolf et al., 2019) and the Adam optimizer. On the Photosynthesis dataset, due to memory requirements, training required a batch size of 8; all other datasets were trained with a batch size 16. The learning rate was tuned individually for each dataset with a grid of {2e-5, 3e-5, 5e-5}. Matching the RNN model, an exponential moving average over the model's weights was employed during training. Hyperparameters were tuned for 20 epochs.

For all experiments, we trained models with 10-fold cross validation with train/validation/test splits, evaluating on pooled (concatenated) predictions across folds. We split the data into 80% train, 10% validation, and 10% test. For hyperparameter tuning, we trained on each train split and evaluated performance on the validation split, retaining

Item	Model	Corr	QWK	MSE	Sig.
MI	SVR	0.7804	0.7045	0.3298	
	RNN	0.7989	0.7642	0.3058	
	PT	**0.8134**	**0.7733**	**0.2956**	
PS	SVR	0.8296	0.7851	0.2098	R
	RNN	0.8215	0.7550	0.2285	
	PT	**0.8459**	**0.8246**	**0.1997**	S,R
SO	SVR	0.7491	0.6690	0.2737	
	RNN	0.7612	0.7116	0.2619	
	PT	**0.7691**	**0.7127**	**0.2608**	S,R
TC	SVR	0.6856	0.6156	0.4777	
	RNN	0.7106	0.6732	0.4465	S
	PT	**0.7286**	**0.6791**	**0.4266**	S

Table 3: Human-machine agreement for Knowledge Integration (KI) score models. QWK = quadratic-weighed kappa, MSE = mean squared error. SVR = support vector regression, RNN = recurrent neural network, PT = pre-trained Transformer. *Sig.* = significance by bootstrap replicability analysis; see main text for details.

the predictions from the best performance across epochs and the epoch on which that performance was observed. We pooled the predictions from all folds on the validation sets, evaluated performance, and selected the best-performing configuration of hyperparameters. For final model training, we trained models on combined train and validation splits, again with 10-fold cross-validation, to the median best epoch across folds from the hyperparameter tuning phase. Final performance was evaluated on the pooled predictions from the test splits. This training and evaluation procedure improves the stability of estimates of performance during both the tuning and final testing phases and makes use of more data for training and evaluating the final models, providing better estimates of model performance.

3.4 Evaluation metrics

To evaluate the agreement of human scores and machine scores, we report Pearson's correlation, quadratic weighted kappa (QWK), and mean squared error (MSE). QWK is a measure of agreement that ranges between 0 and 1 and is motivated by accounting for chance agreement (Fleiss and Cohen, 1973). Correlation and MSE are computed over real-valued model predictions, while QWK is computed over rounded predictions.

Item	Sub-score	Model	Corr	QWK	MSE	Sig.
MI	CCC	SVR	.7008	.6314	.3185	
		RNN	.7685	.7322	.2561	S
		PT	**.7730**	**.7542**	**.2557**	S
MI	DCI	SVR	.7505	.7110	.1261	
		RNN	.7908	.7392	.1088	
		PT	**.8230**	**.7970**	**.0953**	
PS	CCC	SVR	.6992	.6050	.3102	
		RNN	**.7379**	**.7187**	**.2772**	
		PT	.7188	.6607	.2997	S
PS	DCI	SVR	.7410	.6956	.2245	
		RNN	.7795	.7471	.1955	
		PT	**.8044**	**.7701**	**.1826**	
SO	eng	SVR	.6957	.5915	.2684	
		RNN	.7484	.7112	.2503	
		PT	**.7662**	**.7263**	**.2428**	S
SO	sci	SVR	.5789	.4770	.1744	
		RNN	**.6872**	.5408	**.1623**	
		PT	.6480	**.6038**	.1834	
TC	exp	SVR	.5323	.4705	.1926	
		RNN	.5916	.4675	.1724	
		PT	**.6067**	**.5445**	**.1661**	
TC	sci	SVR	.5038	.0000	.2262	
		RNN	.5090	.3897	.1835	S
		PT	**.5303**	**.4182**	**.1779**	S

Table 4: Human-machine agreement for NGSS subscore models. *Sig.* = significance by bootstrap replicability analysis; see main text.

4 Results

4.1 Human-machine agreement

The models for the KI scores showed mostly good agreement with human scores (Table 3). QWK for the Musical Instruments, Photosynthesis, and Solar Ovens items was substantially higher than the standard 0.7 recommended for human-machine agreement in real-word automated scoring applications (Williamson et al., 2012).

For NGSS subscore models (Table 4), those with more balanced score distributions (cf. Figure 1) showed good human-machine agreement, while the models trained on the most skewed data distributions showed lower levels of human-machine agreement. Specifically, Solar Ovens-Science and the Thermodynamics Challenge subscore models were trained on data where about 80% of responses had the lowest score. Each of these models' agreement with the human-scored data was relatively low and significantly below the 0.7 QWK threshold.

Across both KI score models and NGSS subscore models, the pre-trained transformer models showed higher human-machine agreement than both the SVR and RNN models in almost all cases. On the KI score datasets, the performance improve-

ment from the PT models was relatively modest, except for the Photosynthesis dataset, where a larger improvement was observed. On the NGSS subscore datasets, the improvement from the PT models was often larger. This may be the result of stronger representations from the pretrained models compensating from the smaller training dataset sizes. At the same time, RNN models also performed well on data-impoverished datasets such as Photosynthesis-CCC and Solar Ovens-science.

The cross-validation training and evaluation procedure employed here poses a challenge to statistically estimating the strengths of the differences between methods since the folds are not independent. Here we employ replicability analysis for multiple comparisons (Reichart et al., 2018; Dror et al., 2017). We use bootstrap-based significance testing on each fold for the final model on each dataset and then perform *K-Bonferonni* replicability analysis. We define significance as rejecting the null hypothesis of no difference for at least half of the folds. The results of these hypothesis tests are shown in Tables 3 and 4. For example, *S* indicates the model in that row (PT) performed significantly better than the SVR model (similarly for the RNN models). Although this hypothesis testing framework is conservative, the results support the conclusion that the pre-trained transformer models' performance was strong.

5 Error analysis

In this section, we explore the differences in the two neural models (RNN and PT) in more detail by looking at patterns of errors. We focus on instance-level saliency maps – gradient-based methods that identify the importance of tokens to the model by examining the gradient of the loss. For each dataset, we sample 100 responses and generate saliency maps for each. We use the *simple gradient* method (Simonyan et al., 2014) via AllenNLP (Wallace et al., 2019). The item developers manually analyzed the generated saliency maps for each response and model.

We analyzed two sets of cases:

1. One neural model accurately predicted the human score while the other did not. How do the error patterns in these cases illustrate how the models each learned differently from the training data?

2. Both models incorrectly predicted the human score, and moreover predicted the same incorrect score. Do the models make the wrong prediction for the same or different reasons?

In the following, due to space constraints, we focus on error analysis for the scoring model for the Musical Instruments knowledge integration dataset.

One correct, one incorrect. Cases where one model accurately predicted the human score while the other did not illuminated several differences in the two neural models.

The RNN model tended to ignore or de-emphasize some keywords, while overemphasizing high frequency and function words. For example, Figure 2a shows a simple example where the RNN fails to emphasize the keyword *pitch*. The BERT model accurately registers this word as salient, and predicts the correct score. Similarly, in Figure 2b, the RNN misses the keyphrase *full glass* while the BERT model catches it. In Figure 2c, the RNN spuriously treats the function words *when* and *you* as salient and over-predicts the score.

For its part, the BERT model may de-emphasize many high frequency words but at the same time may regard discourse markers as salient. An example is in Figure 3a, where the BERT model emphasizes *because since*, and this may in part help the model reach the correct prediction.

If the BERT model is able to better learn important keywords (while ignoring more function words), it may sometimes "overlearn" the importance of those tokens, leading to over-prediction of scores. There are several examples where the model uses the word piece *##brate* to overpredict a score (Figure 3b).

Both incorrect with the same prediction. In many cases, the models made the same incorrect predictions for different reasons. An example is Figure 3c, where the RNN emphasizes *deeper* and *dense* while the BERT model focuses on *because* and *cup*. Overall, the same differences in the models identified above held for these cases of making the same incorrect prediction.

In general, although there was some variability across models, both models correctly identified the keywords necessary for scoring responses correctly, leading to good human-machine agreement. The RNN model may be more sensitive to tokens that are good indicators of the score in the training data (either high or low) but not in language in general, such as high frequency and function words, while

148053 score=3 prediction=2
the pitch gets a lot lower
148053 score=3 prediction=3
[CLS] the pitch gets a lot lower [SEP]

(a)

207529 score=3 prediction=2
The tap of a full glass is more low pitched and an empty glass is more high pitched because there is no bellow
207529 score=3 prediction=3
[CLS] the tap of a full glass is more low pitched and an empty glass is more high pitched because there is no bell ##ow [SEP]

(b)

147925 score=2 prediction=3
When you tap on a full glass the pitch stays the same as if you were tapping on an empty glass because you are still tapping on a glass that is going to make a high pitched sound no matter if it is full or not .
147925 score=2 prediction=2
[CLS] when you tap on a full glass the pitch stays the same as if you were tapping on an empty glass because you are still tapping on a glass that is going to make a high pitched sound no matter if it is full or not . [SEP]

(c)

Figure 2: Error analysis: RNN model trends. In each example, the RNN model's saliency map appears on *top*.

237142 score=3 prediction=2
The pitch of the tapped full glass is lower than the pitch of the tapped empty glass because since there is water inside you are not going to be able to hear it as much .
237142 score=3 prediction=3
[CLS] the pitch of the tapped full glass is lower than the pitch of the tapped empty glass because since there is water inside you are not going to be able to hear it as much . [SEP]

(a)

148661 score=3 prediction=3
The one taht is full will vibrate less so it will be higher than the one that is empty .
148661 score=3 prediction=4
[CLS] the one ta ##ht is full will vi ##brate less so it will be higher than the one that is empty . [SEP]

(b)

176754 score=4 prediction=3
the cup with water has a deeper sound because its changing through the dense water but the cup with no water stays the same because the sound wave does n't have to go through anything or change anything .
176754 score=4 prediction=3
[CLS] the cup with water has a deeper sound because its changing through the dense water but the cup with no water stays the same because the sound wave doesn ' t have to go through anything or change anything . [SEP]

(c)

Figure 3: Error analysis: Pre-trained transformer model trends. In each example, the pre-trained transformer model's saliency map appears on the *bottom*.

BERT's pre-training regime may equip it to reduce any reliance on such tokens.

Notably, however, while the models usually made good use of keyword evidence to arrive at correct scores, when the models made inaccurate predictions, it was often because the response had the right vocabulary but the wrong science. For example, in the Musical Instruments item, a response might contain *pitch*, *lower*, *density*, and *vibrations*, but the response might attribute the lower pitch to the empty glass. At least two issues were observed in cases of model mis-prediction: (1) students used anaphoric *it* to refer to key concepts (e.g., *full glass* or *empty glass*), but the models do not incorporate anaphora resolution capabilities; (2) models fail to associate the right keywords with the right concepts, in the way that human raters did.

6 Related work

The task of automated content scoring has recently gained more attention (Kumar et al., 2017; Riordan et al., 2017; Burrows et al., 2015; Shermis, 2015). Our work is similar to Mizumoto et al. (2019), who developed a multi-task neural model for assigning an overall holistic score as well as content-based analytic subscores. We leave a multi-task formulation of our application setting for future work.

Sung et al. (2019) demonstrated state-of-the-art performance for similarity-based content scoring on the SemEval benchmark dataset (Dzikovska et al., 2016). In this work, we use pre-trained transformer models for instance-based content scoring (cf. Horbach and Zesch (2019)). That is, we use whole responses as training data and fine-tune pre-trained representations for response tokens on the content score prediction task.

Recently, methods have been introduced to incorporate "saliency" directly into the model training process (Ghaeini et al., 2019). The current work focuses on interpreting the predictions of models trained without additional annotations (for an overview of interpretability in NLP, see Belinkov and Glass (2019). Exploring the contribution of augmented datasets and training algorithms is future work. To our knowledge, our work is the first to to explore the relevance of the saliency in the predictions of neural methods for the content scoring task.

7 Conclusion

We described a set of constructed response items for middle-school science curricula that simultaneously assess students on expression of NGSS Disciplineary Core Ideas (DCIs), Cross-Cutting Concepts (CCCs), and Science and Engineering Practices (SEPs), and the integrative linkages between each, as part of engaging in scientific explanations and argumentation. We demonstrated that human and automated scoring of such CRs for the NGSS dimensions (via independent subscores) and the integration of knowledge (via Knowledge Integration scores) is feasible. We demonstrated that automated scoring can be developed with promising accuracy.

Comparing feature-based, RNN, and pre-trained transformer models on these datasets, we observed that the pre-trained transformer models obtained higher rates of human-machine agreement on most holistic KI score and NGSS subscore datasets. While the RNN models were often competitive with the pre-trained transformer models, an analysis of the different kinds of errors made by each model type indicated that the pre-trained transformer models may be more robust to strong dataset-specific, but spurious, cues to score prediction.

Results showed that, in the formative setting targeted by the online science learning environment used in this study, students often scored at the lowest levels of all three rubrics, which increased skewness in the datasets and likely contributed to reduced model accuracy. Future research will explore more robust methods for learning scoring models from less data in formative settings, especially from highly skewed score distributions, while continuing to provide accurate scoring.

Our findings demonstrate the ability to both develop and automatically score NGSS-aligned CR assessment items. With further refinement, we can provide teachers with both the instructional and technological assistance they need to effectively and efficiently support their students to demonstrate the multidimensional science learning called for by the NGSS.

Acknowledgments

We thank Aoife Cahill for many useful discussions and three anonymous reviewers, Beata Beigman Klebanov, Debanjan Ghosh, and Nitin Madnani for helpful comments. This material is based upon

work supported by the National Science Foundation under Grant No. 1812660. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Rethinking Complex Neural Network Architectures for Document Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Claudia Leacock. 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.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Libby F. Gerard and Marcia C. Linn. 2016. Using Automated Scores of Student Essays to Support Teacher Guidance in Classroom Inquiry. *Journal of Science Teacher Education*, 27(1):111–129.

Reza Ghaeini, Xiaoli Z. Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Saliency Learning: Teaching the Model Where to Pay Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Andrea Horbach and Torsten Zesch. 2019. The influence of variance in learner answers on automatic content scoring. *Frontiers in Education*, 4:28.

Sachin Kumar, Soumen Chakrabarti, and Shourya Roy. 2017. Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Marcia C. Linn and Bat-Sheva Eylon. 2011. *Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration*. Routledge, New York.

Marcia C. Linn, Libby Gerard, Kihyun Ryoo, Kevin McElhaney, Ou Lydia Liu, and Anna N Rafferty. 2014. Computer-guided inquiry to improve science learning. *Science*, 344(6180):155–156.

Ou Lydia Liu, Joseph A. Rios, Michael Heilman, Libby Gerard, and Marcia C. Linn. 2016. Validation of Automated Scoring of Science Assessments. *Journal of Research in Science Teaching*, 53(2):215–233.

Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. 2019. Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring. In *14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA@ACL)*.

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

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

Roi Reichart, Rotem Dror, Gili Baumer, and Segev Shlomov. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Brian Riordan, Michael Flor, and Robert Pugh. 2019. How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA@ACL)*.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating neural architectures for short answer scoring. In *12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA@EMNLP)*.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICLR)*.

Chul Sung, Tejas I. Dhamecha, and Nirmal Mukhi. 2019. Improving Short Answer Grading Using Transformer-Based Pre-training. In *Proceedings of the 20th International Conference on Artificial Intelligence in Education (AIED)*.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matthew Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.