Analyzing automated content scoring for knowledge integration in science explanations using saliency maps

Brian Riordan[1], Sarah Bichler[2], Allison Bradford[2], Marcia C. Linn[2]
[1]ETS
[2]University of California-Berkeley

Abstract

Models for automated scoring of content in educational applications continue to demonstrate improvements in human-machine agreement, but it remains to be demonstrated that the models achieve gains for the "right" reasons. For providing reliable scoring and feedback, both high accuracy and construct coverage are crucial. In this work, we provide an in-depth quantitative and qualitative analysis of automated scoring models for science explanations of middle school students in an online learning environment that leverages *saliency maps* to explore the reasons for individual model score predictions. Our analysis reveals that top-performing models can arrive at the same predictions for very different reasons, and that current model architectures have difficulty detecting ideas in student responses beyond keywords.

1. Introduction

Current machine learning models for automated scoring of content in educational applications have better human-machine agreement than classical machine learning models. While these larger and more complex scoring models often achieve high accuracy, to what extent are the models achieving performance improvements for the right reasons?

Recent work has shown gains in human-machine agreement from neural network models, particularly recurrent neural networks (RNNs) and pre-trained transformer (PT) models. However, prior research has neglected investigating the reasons for improvement at the response level.

In this work, we provide an in-depth quantitative and qualitative analysis of automated scoring models for science explanations of middle school students in an online learning environment that leverages *saliency maps* to explore the reasons for individual model score predictions. Saliency maps provide a visual representation of the importance of each word in a response. They are computed from a model's internal parameters when making score predictions.

Through expert analysis of saliency maps, we focus on the extent to which models attribute importance to words and phrases in student responses that align with item rubrics. We analyze these trends for evidence about how state-of-the-art models carry out the content scoring task in this domain and compare trends by model classes (in particular, RNNs and PT models) to elucidate the differences in top-performing neural models' predictions.

2. Data

2.1. Background

This work focuses on constructed response (CR) items in formative assessments that are embedded in science units for middle school students accessed via an online classroom system (Gerard & Linn, 2016; Linn et al., 2014). The items were scored with a knowledge integration (KI) rubric (Liu 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 rewards students for linking evidence to claims and for adding multiple evidence-claim links to their explanations (Linn & Eylon, 2011).

For this study, we focus in detail on two formative assessment items. KI scoring rubrics and example responses for each item are shown in Table 1.

Musical Instruments and the Physics of sound waves (MI). The Musical Instruments and Physics of Sound Waves unit focuses on developing student ideas about properties of sound waves (wavelength, frequency, amplitude, and pitch). The CR item we designed aligns with the NGSS PE MS-PS4-2 performance expectation and assesses students' understanding of the relationship of pitch and frequency and the characteristics of a sound wave when transmitted through different materials. 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.

Solar Ovens (SO). The Solar Ovens unit asks students to collect evidence and decide whether 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. The embedded CR item assesses how students integrate their ideas about energy transfer and transformation with their interpretations of data about the impact of the solar oven design.

2.2. Data collection

Students from 11 middle schools participated in either a benchmark assessment containing items across several science topics at the beginning or end of the school year or took a pre- and post-test. Across schools, students who received free or reduced-price lunch ranged from 1.6% - 89%, were 50.8% - 97.3% non-white, and were 2.2% - 38.2% English learners.

3. Methods

3.1. Scoring Procedure

For each item, two researchers carried out an iterative process of scoring 10% of the data independently, discussing disagreements, and refining the KI rubric until Cohen’s Kappa reached 0.8. Using the final rubric, one researcher re-scored the entire dataset.

3.2. Models

Modern recurrent neural network models in natural language processing follow a recipe of pre-trained “embeddings” (real-valued vectors from models trained on a different NLP task) to represent words and a model architecture that processes word tokens one at a time. As part of this processing, words’ representations are contextualized by the words in close proximity. The

resulting sequence of vectors can be "pooled" into a single vector with an "attention" mechanism that focuses on parts of each vector in the sequence, or by taking the maximum value in one cell of the vectors across all the vectors in the input.

Pre-trained transformer models leverage word representations that are learned from language models trained on very large corpora. A language model learns to predict words in the input corpus that are "masked out" during training. This "self-training" of predicting words in the input, when coupled with the "transformer" neural network architecture, yields word representations that are useful across many NLP tasks. Unlike RNNs, transformer networks only use a form of attention – self-attention -- between word-like representations.

For this investigation, as a recurrent neural network, we employ a single-layer recurrent neural network model with "maximum pooling" for word vector aggregation. As a pre-trained transformer model, we use a model based on the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin, Chang, Lee, & Toutanova, 2018). Both models have achieved state-of-the-art performance on the ASAP Short Answer Grading dataset (Riordan, Flor, & Pugh, 2019; Steimel & Riordan, 2020). See the Appendix for details about model architecture and training.

The models were trained to predict an ordinal score from each response's text. Since knowledge integration is concerned with the content of the response, the models did not consider grammatical or usage errors in scoring.

3.3. Evaluation

Human-machine agreement. 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 & Cohen, 1973).

Saliency. Our main evaluation focuses on methods for estimating the importance of a word token for a model's score prediction. This "saliency" estimation procedure produces a (normalized) scalar value for each token. Since neural network models are trained by Stochastic Gradient Descent (SGD), we employ gradient-based saliency estimation methods, in which gradient of the model's loss from the error in its score prediction is backpropagated to the model's input level (i.e., the word tokens represented as "embeddings") and aggregated per word token.

We visualize the saliency of each token in each response with "saliency maps" (e.g., Figure 2). For each dataset, we sampled 100 responses and generated saliency maps for each. We used the *simple gradient* method (Simonyan, Vedaldi, & Zisserman (2014) via AllenNLP (Wallace et al., 2019). The item developers manually analyzed the generated saliency maps for each response and model.

To explore trends in saliency according to each type of model, we sampled 25 responses from each of four outcome conditions: both models were correct, one model was correct and the other incorrect (and vice versa), and both models were incorrect (Table 2). Since PT models often perform better than RNN models, we are interested in explaining performance gains with trends from the RNN-,PT+ condition. At the same time, trends in model saliency scores may prove to derive from model behavior that is shared across all outcome conditions.

To analyze responses from each outcome condition with a common framework, each sampled response was labeled by an item developer with one or more categories that represented hypotheses about what tokens the model used to make a prediction, as evidenced by the saliency scores (Table 3). The set of categories was designed to be general enough to apply to any item's data.

4. Results

4.1. Human-machine agreement

The human-machine agreement for each item is displayed in Table 4. The PT model performs slightly better than the RNN model on both items.

4.2. Distribution of outcome conditions

Table 5 and Table 6 show the number and percentage of cases for each outcome condition for each item. First, the percentage of cases of RNN+,PT- and RNN-,PT+ are very similar within each item's results, underlining the competitive performance of the model types. Second, the percentage of RNN-,PT- is more than double the cases where one model was incorrect (i.e., RNN+,PT- and RNN-,PT+), and this percentage was similar across items. Figure 1 presents these trends visually. These results indicate a moderate level of similarity of patterns of predictions: the model types share many more cases where they both make incorrect predictions than cases where individual model types are incorrect.

4.3. Saliency

Comparing the RNN and PT models by saliency label across all outcome conditions (Table 7), some trends emerged on the MI item: the RNN missed links between keywords somewhat more than the PT model (35 vs. 28), and the RNN had slightly higher numbers of cases of *non-keyword is salient* and *did not consider the context of keywords.* On the SO item, the differences between model types across the saliency labels was smaller and hence the trends more uncertain.

Table 8 provides the detailed distribution of saliency labels by model and outcome condition. We highlight several trends. First, the number of examples of *Captured the most important keywords* is similar across model types for the MI item. For the SO item, when models were wrong (i.e. outcome type = RNN+,PT- and model = PT; and outcome type = RNN-,PT+ and model = RNN), the models were less likely to identify the important keywords (outcome type = RNN+,PT-: PT 17, RNN 24; outcome type = RNN- ,PT+: PT 24, RNN 18). Second, on the MI item, marking non-keywords as salient was an issue when models were wrong (outcome type = RNN+,PT-: PT 9, RNN 14; outcome type = RNN-,PT+: PT 3, RNN 10). This was not the case on the SO item. Third, not considering the context of keywords was a particular problem when both models were wrong (item = MI, outcome type = RNN-,PT-: PT 9, RNN 12; model = SO, outcome type = RNN- PT-: PT 14, RNN 12). Moreover, on the SO item in particular, when one model was wrong, it was far more likely to not have considered context (outcome type = RNN+,PT-: PT 12, RNN 1; outcome type = RNN-,PT+: PT 2, RNN 11).

We carried out a detailed qualitative analysis of model behavior based on the saliency labels. First, we examined the trends in each model's errors to discern patterns that might explain the PT model's advantage. Next, we broadened our analysis to consider model behavior in all four

outcome conditions (i.e., both when the models were correct as well as incorrect) to look for differences in saliency that spanned all types of responses. Due to space, we focus on the Musical Instruments (MI) item.

Figure 2 shows several examples of saliency maps for RNN and PT model errors. One noticeable trend for RNN errors was attaching saliency to high frequency or function words. On response 230094, the model marked *an, would, can,* and *depending,* and on 188198, the tokens *and* and *is* were salient. The trends for PT model errors were more subtle and heterogeneous. The PT model sometimes registered more general words (non-keywords) as salient, but typically avoided high frequency function words. The PT model sometimes marked discourse connectives such as *because* (response 190674).

Across outcome conditions, the patterns of salience are often substantially different between RNN models and PT models. These different patterns, however, can still result in the same model predictions (Figure 3). On one hand, the models can make the same *correct* predictions but with different saliency profiles. On response 191704, the RNN and PT models agreed on the salience of *lower*, but differed greatly in the importance of the key phrases *full glass* and *more mass*. On response 190386, the models differed even more, with different levels of salience attached to most words. At the same time, the models can make the *same incorrect* predictions with different saliency profiles. Response 148006 is a simple example: the RNN emphasized *lower*, while the PT emphasized *glass* -- but both models made the same significant over-prediction of the score.

From our analysis, the different patterns in saliency across models do not seem to indicate greatly differing model capabilities. First, the model errors attributable to a lack of consideration of word context provide examples of the models identifying the right keywords but the wrong science, which in turn leads to over-prediction of scores. Response 190019 in Figure 4 is an example. In this item, the phrase *waves flow quicker* is correct when referring to the full glass, but not the empty glass, which this student refers to. Moreover, the response indicates the inaccurate idea that sound travels from one side of the glass to the other if there is nothing (i.e., no water) blocking it. The models seem to accumulate the simple key phrases such as *time*, *filled glass*, *empty glass*, and *waves flow quicker* to predict a higher score than the actual ideas in the response warrant.

Second, the models can identify the right keywords but then not associate those keywords with the correct score. Figure 5 shows an example of both models under-predicting the score of a response. In this case, *reverberate* only appears in the training data once and is associated with a low score (2) (because that response had other deficiencies). As a result, the models likely simply associated *reverberate* with incorrectness and used it as a "short cut" to predicting a lower score than was actually warranted.

5. Conclusion

This work reports on a quantitative and qualitative investigation of how state-of-the-art short answer scoring models make predictions by analyzing the saliency that the models attribute to parts of student responses in the prediction process. Our analysis shows that different classes of state-of-the-art machine learning models for short answer scoring can produce substantially different "saliency profiles" while often predicting the same scores for the same student

responses. While there is some indication that PT models are better able to avoid spurious correlations of high frequency words with scores, overall the models do not seem to differ greatly in their "basic intelligence" – for example, learning statistical correlations between individual words and scores, rather than between *ideas* and scores. These results suggest the need for strategies to build models with natural language understanding capabilities that better represent the constructs targeted by short answer science assessment items.

Acknowledgements

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT),* 4171--4186.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619.

Gerard, L. F., & Linn, M. C. (2016). Using Automated Scores of Student Essays to Support Instructor Guidance in Classroom Inquiry. *Journal of Science Instructor Education*, *27*(1), 111–129.

Linn, M. C., & Eylon, B.-S. (2011). *Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration*. New York: Routledge.

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science*, *344*(6180), 155–156.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of Automated Scoring of Science Assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233.

Riordan, B., Flor, M., & Pugh, R. (2019). How to account for *mispellings*: Quantifying the benefit of character representations in neural content scoring models. *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA@ACL)*, 116–126.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *International Conference on Learning Representations (ICLR)*.

Steimel, K., & Riordan, B. (2020). Towards Instance-Based Content Scoring with Pre-Trained Transformer Models. *Workshop on Artificial Intelligence for Education (AI4EDU@AAAI)*.

Wallace, E., Tuyls, J., Wang, J., Subramanian, S., Gardner, M., & Singh, S. (2019). AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, *abs/1910.03771*.

Appendix

RNN model. GloVe 100 dimension pre-trained embeddings were used to embed word tokens and fine-tuned during training. Out-of-vocabulary tokens were mapped to a randomly initialized UNK embedding. The supervision of the network was response scores scaled to [0, 1] prior to training. For evaluation, the scaled scores were converted back to their original range. The recurrent neural network (RNN) models used Gated Recurrent Units in 1 layer with a 250-dimension hidden state. The training objective was minimization of a mean squared error loss function. The RNN was optimized with RMSProp with *rho* of 0.9, learning rate 0.001, batch size 32, and gradient clipping (10.0). An exponential moving average was used to smooth the model's weights across training epochs (decay rate = 0.999). In the hyperparameter tuning phase, models were trained for 50 epochs.

PT model. The pre-trained transformer (PT) model used the "bert-base-uncased" pre-trained instance (Wolf et al., 2019). The model was optimized with Adam, a learning rate tuned from {2e-5, 3e-5, 5e-5}, batch size 16, and an exponential moving average of the model weights. Hyperparameters were tuned for 20 epochs.

Model training. Models were trained with 10-fold cross validation with train/validation/test (80/10/10) splits. Predictions were pooled (concatenated) across folds and used for evaluation. For hyperparameter tuning, we trained on each train split and evaluated performance on the validation split, keeping the best predictions across epochs and the epoch on which that performance was observed. Specifically, predictions were pooled from all folds on the validation sets, performance was evaluated, and the best-performing configuration of hyperparameters was selected. For final model training, models were trained on combined train and validation splits 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 aims to increase the stability of estimates of performance during both the tuning and final testing phases and to use more data for training and evaluating the final models in order to provide better estimates of model performance.

Table 1. Knowledge integration (KI) scoring rubrics and example responses.

Score	Description		
		Solar Ovens	Musical Instruments
1	Off-task	David's claim is ... because ...idk	it is just how it works
2	On-task but lacks normative ideas	he is correct because when you look on how fast it heats up mostly all the heat energy was there.	It will always stay the same because the spoon is the same
3	Partial link - normative ideas without any valid links between normative ideas	David's claim is wrong because the wide short was a bigger target for the sunrays to hit so more heat got into the box.	The pitch is lowered by the water in the glass. This means that the glass full of water will have a lower pitch than the glass that is empty.
4	Full link - one valid link between normative ideas	David's claim is was completely wrong because the skinny long box opening was too small not allowing sun light to go inside. That why its better to use the wide box because it has more of a bigger window for the sun light to in.	They will be different because when you had a more dense medium like water into a cup instead of less dense air the sound gets caught more between the particles resulting in a lower pitch.
5	Complex link - multiple valid links between normative ideas	David's claim is incorrect because based on the information I collected form the computer model, the short and wide increased its temperature. The movement of energy causes one solar oven to heat up faster than the other because the wide opening gap lets the infrared radiation, in the inside, becomes heat.	I think that the glass full of water would have a lower pitch because the cup would have more mass which would make the cup harder to vibrate which makes the sound so it would have a lower pitch. The sound waves would also have a longer wavelength and would have a lower frequency.

Table 2: Outcome distribution.

	RNN correct	PT correct
RNN+,PT+	+	+
RNN+,PT-	+	-
RNN-,PT+	-	+
RNN-,PT-	-	-

Table 3: Labels for model saliency behavior.

Captured the most important keywords	Key words that indicate correct relationships, i.e., accurate understanding, are marked as salient.
Missed link between keywords	The model highlights key words but does not attach salience some key words that together with the ones highlighted lead to a score decision.
Non-keyword is salient	Words that are not indicative for accurate understanding are salient.
Did not consider context of keywords	Key words that usually indicate accurate understanding were recognized but the context (presence of other key words) are missed. In the context of other key words, the identified key words do not indicate accurate understanding.

Table 4. Human-machine agreement.

Item	Model	Pearson	QWK	MSE
Musical Instruments	RNN	0.7989	0.7642	0.3058
	PT	0.8134	0.7733	0.2956
Solar Ovens	RNN	0.7612	0.7116	0.2619
	PT	0.7691	0.7127	0.2608

Table 5: Musical Instruments outcome distribution.

	PT correct	**PT incorrect**
RNN correct	762 (0.583)	138 (0.106)
RNN incorrect	132 (0.101)	274 (0.210)

Table 6: Solar Ovens outcome distribution.

	PT correct	**PT incorrect**
RNN correct	1097 (0.630)	135 (0.078)
RNN incorrect	131 (0.075)	377 (0.217)

Figure 1: Outcome distribution excluding both models correct, normalized.

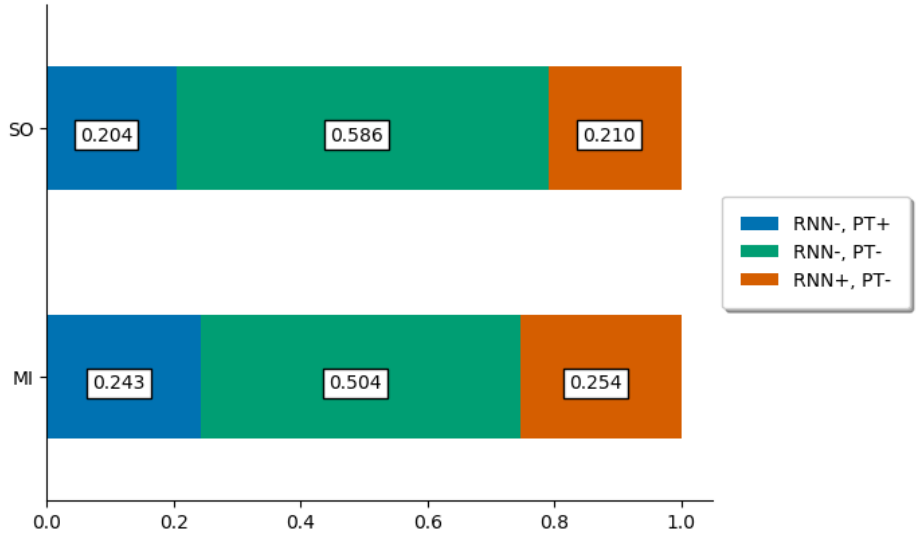

Table 7: Distribution of labels for saliency behavior by item.

Item	Model type	Captured the most important keywords	Missed link between keywords	Non-keyword is salient	Did not consider context of keywords
MI	PT	73	28	32	16
	RNN	73	35	38	21
SO	PT	79	10	45	29
	RNN	83	8	42	25

Table 8: Distribution of labels for saliency behavior by item, model, and outcome condition.

Item	Outcome condition	Model type	Captured the most important keywords	Missed link between keywords	Non-keyword is salient	Did not consider context of keywords
MI	RNN+ PT+	PT	19	10	12	2
	RNN+ PT+	RNN	20	12	4	0
	RNN+ PT-	PT	19	6	9	4
	RNN+ PT-	RNN	19	9	14	6
	RNN- PT+	PT	23	9	3	1
	RNN- PT+	RNN	21	9	10	3
	RNN- PT-	PT	12	3	8	9
	RNN- PT-	RNN	13	5	10	12
SO	RNN+ PT+	PT	22	1	5	1
	RNN+ PT+	RNN	25	0	0	1
	RNN+ PT-	PT	17	3	16	12
	RNN+ PT-	RNN	24	0	14	1
	RNN- PT+	PT	24	0	9	2
	RNN- PT+	RNN	18	5	11	11
	RNN- PT-	PT	16	6	15	14
	RNN- PT-	RNN	16	3	17	12

Figure 2. Examples of errors from attributing saliency to non-keywords for (a) RNN model (b) PT model.

230094 RNN score=3 prediction=2
An empty glass would make one sound but a full glass can make different sound depending on how full the glass is like for example the glass can make different pitches .
188198 RNN score=3 prediction=2
it 's different because one is full and the other is empty .

(a)

190674 PT score=2 prediction=1
[CLS] because there is nothing to block the sound wave for the empty cup of water it i ' ll go faster [SEP]
233477 PT score=3 prediction=3
[CLS] i chose this answer because the empty glass will have a higher pitch sound because the glass is empty . [SEP]

(b)

Figure 3. Different patterns of salience result in the same model predictions.

191704
RNN score=4 prediction=4
If the full glass has more mass in it then the pitch will be lower .
PT score=4 prediction=4
[CLS] if the full glass has more mass in it then the pitch will be lower . [SEP]

(a)

190386
RNN score=3 prediction=3
It is different because the water will slow down the sounds . The more full will make the sound lower .
PT score=3 prediction=3
[CLS] it is different because the water will slow down the sounds . the more full will make the sound lower . [SEP]

(b)

148006
RNN score=1 prediction=3
The glass is lower .
PT score=1 prediction=3
[CLS] the glass is lower . [SEP]

(c)

Figure 4. Both model types showed evidence of a lack of consideration of the context of keywords and phrases.

190019
RNN score=2 prediction=3
This is because the filled glass will take a longer time to travel to the other side than the empty glass . This is because waves flow quicker when there is nothing in their way .
PT score=2 prediction=3
[CLS] this is because the filled glass will take a longer time to travel to the other side than the empty glass . this is because waves flow quicker when there is nothing in their way . [SEP]

Figure 5. Both model types can associate correct keywords with an incorrect score. *reverberate* is associated with a low score in a single response in the training data, which likely leads the models to under-predict the score of the response.

254470
RNN score=4 prediction=3
the empty glass is able to reverberate more and make a high pitch noise .
PT score=4 prediction=2
[CLS] the empty glass is able to rev ##er ##ber ##ate more and make a high pitch noise . [SEP]