Analyzing saliency in neural models for scoring content in science explanations

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 connecting scoring decisions to scoring rubrics are crucial. We provide a 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

Recent work on scoring content in education has shown gains in human-machine agreement from neural network models, particularly recurrent neural networks (RNNs) and pre-trained transformer (PT) models (Mizumoto et al., 2019; Riordan et al., 2019; Sung et al., 2019). However, prior research has neglected investigating the reasons for improvement at the response level. Through expert analysis of saliency maps (Simonyan et al., 2014), we focus on the extent to which models attribute importance to words and phrases in student responses that align with question rubrics. We analyze these trends for evidence about how state-of-the-art models carry out the content scoring task in this domain.

This work focuses on formative assessment questions that are embedded in science units for middle school students accessed via an online classroom system (Gerard and Linn, 2016; Linn et al., 2014). For this study, we focus on two formative assessment questions: (1) Musical Instruments (MI): Students develop ideas about properties of sound waves (wavelength, frequency, amplitude, and pitch). (2) Solar Ovens (SO): Students 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 from 11 U.S. middle schools participated by engaging in the science units during science classes.

2 Methods

RNN and PT models were trained to predict an ordinal score from each response's text. The RNN model was a 1-layer GRU with 250-dimensional hidden state and GloVe 100-dimension embeddings. The model was trained to minimize a mean squared error loss for 50 epochs. The PT model used a `bert-base-uncased` pre-trained instance (Wolf et al., 2019) optimized with Adam and a learning rate tuned from {2e-5, 3e-5, 5e-5} for a maximum of 20 epochs.

Our main evaluation focuses on methods for estimating the importance of a word token for a model's score prediction. We employ gradient-based saliency estimation methods to produce a (normalized) scalar value for each token and visualize the saliency of tokens with "saliency maps" (Figure 1). For each dataset, we sampled 100 responses and generated saliency maps for each response. We used the simple gradient method (Simonyan et al., 2014; Wallace et al., 2019).

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 (i.e. RNN+,PT- and vice versa), and both models were incorrect. We carried out two sets of analyses: First, to analyze responses from each outcome condition with a common framework, each sampled response was labeled by a question developer with

Question	Model type	Pearson	QWK	MSE
SO	RNN	0.7612	0.7116	0.2619
SO	PT	0.7691	0.7127	0.2608
MI	RNN	0.7989	0.7642	0.3058
MI	PT	0.8134	0.7733	0.2956

Table 1: Human-machine agreement. *Pearson* = Pearson's *r*, QWK = quadratically-weighted kappa, *MSE* = mean squared error.

one or more categories that represented hypotheses about what tokens the model used to make a prediction, as evidenced by the saliency scores. The categories were *Captured the most important keywords*, *Missed link between keywords*, *Non-keyword is salient*, and *Did not consider context of keywords*. The set of categories was designed to be general enough to apply to any question's data. Second, we carried out a detailed qualitative analysis of model behavior based on the saliency labels.

3 Results and Discussion

Two important trends in the distribution of saliency labels were: (1) The number of examples of *Captured the most important keywords* was similar across model types for the MI question, but for the SO question, when models were wrong, they were less likely to identify the important keywords (RNN+ PT-: PT 17, RNN 24; RNN- PT+: PT 24, RNN 18). (2) Not considering the context of keywords was a particular problem when both models were wrong (MI RNN- PT-: PT 9, RNN 12; SO RNN- PT-: PT 14, RNN 12). Moreover, on the SO question, when one model was wrong, it was more likely to have ignored context (RNN+ PT-: PT 12, RNN 1; RNN- PT+: PT 2, RNN 11).

We report qualitative analyses on the MI question (Figure 1). Across outcome conditions, the patterns of salience were often substantially different between RNN models and PT models. These different patterns, however, could still result in the same model predictions (responses 191704, 190386). On one hand, the models could 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*. At the same time, the models made the same incorrect predictions with different saliency profiles (response 148006).

From our analysis, the different patterns in saliency across models do not seem to indicate

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]

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]

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

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]

Figure 1: Examples of saliency patterns in RNN and pretrained transformer (PT) model errors.

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. Second, the models can identify the right keywords but not associate them with the correct score – for example, because limited training data creates associations between a score and a rare word. Response 254470 shows an example of both models under-predicting the score of a response because *reverberate* only appears in the training data once and is associated with a lower score.

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, our results indicate that both models focus on learning statistical correlations between scores and words and do not demonstrate an ability to learn key phrases or longer linguistic units corresponding to ideas, which are targeted by question rubrics. These results point to a need for models to better capture student ideas in science assessments.

Acknowledgments

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

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.

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.

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)*.

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)*.

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)*.

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.