Probing Saliency in Short Answer Scoring Models for Science Explanations

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

Summary

Goal: Do pretrained transformer (PT) and RNN models achieve performance gains in short answer scoring for the "right" reasons?

Task: Ordinal score prediction for

Methods: PT- and RNN-based text regression models; expert analysis of saliency maps for responses

Data: U.S. middle school students using an online science platform

Results

- PT- and RNN-based models can produce **substantially different saliency profiles while predicting the same scores** for the same student responses
- Models **do not show an ability to learn key phrases or longer linguistic units corresponding to ideas**, which are targeted by question rubrics

Introduction

- Online science education environment: WISE https://wise.berkeley.edu/
- Knowledge integration (KI) scores: 1-5; reward linking evidence to claims and adding multiple evidence-claim links to explanations

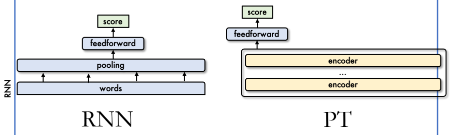
Datasets

Musical Instruments (MI): Students develop ideas about properties of sound waves (wavelength, frequency, amplitude, pitch).

Solar Ovens (SO): Students collect evidence related to a claim made by a fictional peer about the functioning of a solar oven.

Experiment Setup

Architectures



Training procedure

- 10-fold cross-validation (train/dev/test)
- Analysis of "pooled" predictions on test

Analysis procedure

- Sampled 25 responses from 4 outcome conditions: RNN+ PT+ (RNN correct, PT correct); RNN+ PT-; etc.
- Experts viewed the model's saliency map for each response and labeled the model's behavior with 1 or more labels

Labels for Model Saliency Behavior

Captured the most important keywords (+kw)	Key words that are indicative of accurate understanding are salient.
Missed link between keywords (-link)	Some key words are salient but not others; all are required for a credible score decision.
Non-keyword is salient (+nkw)	Some words that are *not* indicative of accurate understanding are salient.
Did not consider context of keywords (-ctxt)	Some typical key words are salient, but in the context of other key words, the identified key words do not indicate accurate understanding.

Quantitative Results

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

Q	Cond.	Model	+kw	-link	+nkw	-ctxt
MI	RNN+	PT	19	10	12	2
	PT+	RNN	20	12	4	0
	RNN+	PT	19	6	9	4
	PT-	RNN	19	9	14	6
	RNN-	PT	23	9	3	1
	PT+	RNN	21	9	10	3
	RNN-	PT	12	3	8	9
	PT-	RNN	13	5	10	12
SO	RNN+	PT	22	1	5	1
	PT+	RNN	25	0	0	1
	RNN+	PT	17	3	16	12
	PT-	RNN	24	0	14	1
	RNN-	PT	24	0	9	2
	PT+	RNN	18	5	11	11
	RNN-	PT	16	6	15	14
	PT-	RNN	16	3	17	12

When a model was wrong, it was less likely to identify the important keyword.

When both models were wrong, not considering the context of keywords was a particular problem.

Qualitative Results

Different patterns of salience sometimes 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]

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]

Both model types sometimes associate correct keywords with an incorrect score.

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]

Both model types sometimes attribute saliency to non-keywords.

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 .

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]

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.