Towards Instance-Based Content Scoring with Pre-Trained Transformer Models

Kenneth Steimel[1]* **and Brian Riordan**[2]
[1]Indiana University
[2]ETS

Abstract

Pretrained contextual word representations based on transformer models have recently achieved state-of-the-art performance on content scoring for educational data using a similarity-based scoring approach with reference answers. In this work, we demonstrate how similar models can be adapted for content scoring using an instance-based approach (Horbach and Zesch 2019), in which a model is learned only from student responses (not reference answers). Our approach yields state-of-the-art performance on the ASAP-SAS short answer scoring dataset.

Content scoring is the task of scoring the content of answers to free-response questions in educational applications (also known as *short answer grading* or *scoring* when responses are short, e.g. sentence length) (Burrows, Gurevych, and Stein 2015). Unlike systems for essay scoring, which target writing quality (e.g., ideas and elaboration, organization, style, and writing conventions such as grammar and spelling (Burstein, Tetreault, and Madnani 2013)), systems for content scoring focus on the accuracy of responses.

Two main approaches exist for content scoring: similarity-based and instance-based (Horbach and Zesch 2019). Similarity-based approaches score content by comparing individual responses with reference responses, while instance-based approaches learn a model of the characteristics of responses at different scores. Recent work on neural methods for content scoring have shown strong performance in both similarity-based (Kumar, Chakrabarti, and Roy 2017) and instance-based (Riordan, Flor, and Pugh 2019) scoring scenarios.

Leveraging recent advances in pre-trained contextual word representations, Sung, Dhamecha, and Mukhi (2019) demonstrated state-of-the-art performance for similarity-based content scoring. Since pretrained models such as BERT are trained on tasks that use intra-sentence and neighboring sentence information, they fit the similarity-based scoring approach well.

In this work, we explore how to apply pretrained transformer models for instance-based content scoring. That is, we use whole responses as training data and fine-tune pretrained representations for response tokens on the content score prediction task. Specifically, in this work, we demonstrate that a BERT-based model can be adapted for instance-based content scoring and examine its performance relative to the state-of-the-art for instance-based content scoring.

*Work carried out during an internship at ETS.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Related Work

Sung et al. (2019) applied a pretrained BERT model to content scoring by formulating the scoring task as sentence pair classification: given a student response and a reference answer, predict the response's classification (correct, incorrect, contradictory). This setup mirrors work on natural language inference and related tasks (Wang et al. 2018), where pretrained contextual word representations have demonstrated state-of-the-art performance. By contrast, we formulate the scoring task as a text regression task similar to sentiment analysis: given a document (response), predict a (integer-valued) score.

Recent work has demonstrated the effectiveness of simple pooling mechanisms for neural models for text classification (Shen et al. 2018; Adhikari et al. 2019) and content scoring (Riordan, Flor, and Pugh 2019). Liu et al. (2019) showed that transformer-based contextualizers like BERT perform better using a scalar mix of different layers, motivating our approach.

Dataset

The Automated Student Assessment Prize Short Answer Scoring (ASAP-SAS) dataset is one of only a few public benchmark short answer scoring datasets. The dataset is made up of 10 questions on academic subjects including science, biology, and English Language Arts. Each question was administered to United States high school students as part of state-level assessments. All responses were scored by two human scorers using scales of 0-2 or 0-3 depending on the question (Shermis 2015). We used the official training and test data.[1] Table 1 provides the mean number of words per prompt for the training and test sets.

[1]`https://www.kaggle.com/c/asap-sas`. The official test set is `public_leaderboard_rel_2.tsv`.

	1	2	3	4	5	6	7	8	9	10	Mean QWK	$\text{Mean}_{\text{Fisher}}$ QWK	Mean MSE
Words (train)	52.50	66.86	54.58	47.07	29.19	27.07	47.26	62.06	56.04	48.42			
Words (test)	54.32	67.37	55.37	47.48	29.24	28.73	48.00	61.62	55.68	46.77			
RNN-based	.8301	.7913	.6620	**.7310**	**.8441**	**.8610**	.7362	.6641	.8087	.7766	.7705	.7788	.2200
BERT-based	**.8533**	**.8430**	**.6988**	.6134	.8232	.8464	**.7577**	**.6742**	**.8518**	**.7795**	**.7741**	**.7858**	**.2055**

Table 1: Numbers of words per response in the train and test sets, and per-prompt quadratic weighted kappa (QWK) and mean results on the official test set.

Pooling Method	Moving average	BERT Layers	Mean MSE
Mean	+	Top	**0.2109**
		Mix	0.2152
	-	Top	0.2153
		Mix	0.2162
Max	+	Top	0.2191
		Mix	0.2177
	-	Top	0.2138
		Mix	0.2132
'CLS'	+	Top	0.2228
		Mix	0.2239
	-	Top	0.2248
		Mix	0.2267

Table 2: BERT-based model performance on validation set.

Method

Network architecture

BERT is a 12-layer bidirectional transformer model trained on the tasks of masked token prediction and next sentence prediction across very large corpora. 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. To use the pretrained model in downstream tasks, the model is 'fine-tuned' by training the model weights directly on the task of interest, making predictions via the '[CLS]' token.

For instance-based content-scoring, we (1) adapt the pooling mechanism across token representations and (2) use representations from all model layers for prediction.

First, instead of using the trained '[CLS]' token and its subsequent pooling layer, we explore *mean* and *max* pooling across both the '[CLS]' token representation and the representations of all wordpieces in the response. This approach more directly uses information available across the response. The result of the pooling layer is a single vector. Since we target integer-valued content scoring, we add a feedforward layer with sigmoid activation and scale scores to [0,1].

Second, we investigate *scalar mixing* across the layers of the model. From the activations for each layer, a vector of weights is learned corresponding to the contribution of each layer. This weighted combination of layers is processed by the pooling mechanism.

Data preparation and model training

Prior to training, the text was spell-corrected with a high-performing system similar to Flor (2012) and Flor and Futagi (2012). All scores of responses were scaled to [0, 1], and these scaled scores were used as input to the networks. For evaluation, the scaled scores were converted back to their original range (Taghipour and Ng 2016). Networks were trained with a mean squared error loss.

We trained models with 5-fold cross validation with train/validation/test splits. We split the official ASAP-SAS training data into 5 folds of 80% train and 20% validation. For hyperparameter tuning, we evaluated performance only on the validation sets and recorded the best performance across epochs. For training final models after hyperparameter tuning, we combined the training and validation sets and stopped training at the average best epoch across validation folds rounded to the nearest epoch (cf. Johnson & Zhang (2017)).

For all experiments, we used a batch size of 16 and tuned the learning rate in {2e-5, 3e-5, 5e-5}. The best results were obtained with a learning rate of 5e-5. We also applied an exponential moving average across model parameters, using a decay rate of 0.999 to match the setting used for the RNN-based model (Riordan, Flor, and Pugh 2019).

Results and Discussion

We report model performance on each prompt with quadratic weighted kappa (QWK). To summarize performance across prompts, we report mean QWK, Fisher-weighted mean QWK (the official metric of the ASAP-SAS competition), and mean MSE.

The mean MSE of model variants on the validation set is shown in Table 3. The mean and max pooling mechanisms produce much lower MSE than the standard BERT pooling mechanism. Taking the scalar mix of layers resulted in mixed performance, yielding lower mean squared error values for max pooling but higher values for mean pooling and standard BERT pooling. Taking the moving average of weights across epochs was often helpful, and contributed to the best configuration. We used mean pooling, only the top layer, and a moving average of the weights for training final models on the test set.

Table 1 shows the per-prompt performance and mean performance across prompts on the test set for both the BERT-based system and the state-of-the-art RNN-based system from Riordan, Flor, & Pugh (2019). The fine-tuned BERT-based system surpasses the state-of-the-art RNN model on

	1	2	3	4	5	6	7	8	9	10	Mean QWK	$\text{Mean}_{\text{Fisher}}$ QWK	Mean MSE
RNN-based	.8379	.7452	.6681	**.6754**	.7856	.8392	.7486	.6740	.7766	.7535	.7504	.7568	.2263
BERT-based	**.8594**	**.7785**	**.7075**	.6520	**.8136**	**.8457**	**.7763**	**.6869**	**.7945**	**.7663**	**.7681**	**.7757**	**.2109**

Table 3: Per-prompt mean QWK across cross-validation fold dev sets.

the QWK metrics. While the two systems' performance is similar on most prompts, there are significant differences (prompts 2, 4, and 9). While investigating the robustness of these differences is the subject of ongoing work, we conjecture that the differences are largely due to suboptimal selection of test epoch on which to measure performance (based on the average best epoch from cross-validation on the training set). Indeed, when considering the mean performance per prompt across validation sets from cross-validation folds (Table 3), the difference in performance between the model types becomes clearer: the BERT-based model's performance is overall stronger.

In this work, we explored how pretrained transformer models can be successfully adapted to instance-based content scoring. We achieved a new state-of-the-art human-machine agreement score on the ASAP-SAS dataset. We investigated simple pooling mechanisms, scalar mixing of layers, and moving averages of model weights during the fine-tuning process. Future work will analyze the representations of pretrained transformer models for their contributions to scoring performance.

References

Adhikari, A.; Ram, A.; Tang, R.; and Lin, J. 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, Volume 1 (Long and Short Papers)*, 4046–4051. Minneapolis, Minnesota: Association for Computational Linguistics.

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.

Burstein, J.; Tetreault, J.; and Madnani, N. 2013. The e-rater automated essay scoring system. *Handbook of automated essay evaluation: Current applications and new directions* 55–67.

Flor, M., and Futagi, Y. 2012. On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Flor, M. 2012. Four types of context for automatic spelling correction. *Traitement Automatique des Langues (TAL)* 53(3):61–99.

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

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)*.

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

Liu, N. F.; Gardner, M.; Belinkov, Y.; Peters, M.; and Smith, N. A. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Riordan, B.; Flor, M.; and Pugh, R. 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)*.

Shen, D.; Wang, G.; Wang, W.; Min, M. R.; Su, Q.; Zhang, Y.; Li, C.; Henao, R.; and Carin, L. 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)*, 440–450. Melbourne, Australia: Association for Computational Linguistics.

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

Sung, C.; Dhamecha, T. I.; and Mukhi, N. 2019. *Artificial Intelligence in Education*. Springer. chapter Improving Short Answer Grading Using Transformer-Based Pre-training, 469 – 481.

Taghipour, K., and Ng, H. T. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics.