2023 IEEE International Conference on Advanced Learning Technologies (ICALT) | 979-8-3503-0054-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICALT58122.2023.00104

Neural Automated Short-Answer Grading Considering Examinee-Specific Features

Masaki Uto
The University of Electro-Communications
Chofu, Tokyo
uto@ai.lab.uec.ac.jp

***Abstract*—Automated short-answer grading (ASAG) is the task of automatically assigning scores to examinees' textual responses to short-answer questions. Recently, various ASAG models based on deep neural networks (DNNs) have been proposed and some have achieved high accuracy. Conventional ASAG models are generally trained and used independently for each short-answer question, even when a given test contains multiple short-answer questions. However, because tests are tools for evaluating particular examinee traits, multiple questions on the same test are generally designed to measure similar latent traits in examinees. Latent examinee traits across multiple short-answer questions may thus function as effective auxiliary features for ASAG. We therefore propose a new DNN-based ASAG model with an examinee-aware architecture that extracts examinee-specific features, including latent traits, from answer texts across multiple questions.**

***Index Terms*—Automated short-answer grading, educational assessment, deep neural networks**

I. Introduction

Short-answer questions are widely used in various educational tests. However, the use of short-answer questions, especially in large-scale tests, has prompted concerns related to scoring reliability, time complexity, and monetary cost. Automated short-answer grading (ASAG) has attracted attention as a way to alleviate these concerns [1]. Conventional ASAG models have relied on manually designed features, while more recent ASAG models are designed based on deep neural networks (DNNs) (e.g., [1]–[8]). DNN-based ASAG models can automatically extract latent features that are effective for score prediction through model training using a dataset of scored short-answer texts, and some have achieved high scoring accuracy.

Conventional DNN-based ASAG models are generally trained and used independently for each short-answer question, even when a given test contains multiple short-answer questions. Therefore, they cannot consider any shared factors that affect examinees' performance on multiple questions, even if such factors are assumed to exist. Meanwhile, because tests are tools for evaluating particular examinee traits, multiple questions on the same test are generally designed to measure similar latent traits in examinees [6], [7]. Such latent examinee traits across multiple short-answer questions may thus function as effective auxiliary shared features for ASAG.

We therefore propose a new DNN-based ASAG model with an examinee-aware architecture consisting of neural layers shared across multiple questions, which extracts examinee-specific latent features including latent traits. Specifically, the proposed model is formulated as a multi-input and multi-output DNN model that receives an examinee's multiple short-answer texts for multiple questions from a single test and simultaneously outputs multiple corresponding scores. The proposed model first extracts distributed representations of answer texts through a text encoder, for which various conventional DNN-based ASAG models are applicable. Then, the examinee-aware architecture receives distributed representations of multiple answer texts and extracts examinee-specific latent features, which are expected to reflect latent traits, using a concatenated vector of those distributed representations. Examinee-specific features and text-specific distributed representations are then used to predict scores for input texts. This paper demonstrates the effectiveness of the proposed model through experiments using actual data.

II. Neural Automated Short-Answer Grading

This section introduces representative conventional DNN-based ASAG models using a recurrent neural network (RNN) and Bidirectional Encoder Representations from Transformers (BERT).

The RNN-based ASAG model [2] receives a word sequence in a short-answer text as input and then outputs a score through 1) a word embedding layer, which transforms each word to a word-embedding representation, 2) an RNN layer, which transforms each word-embedding vector to another vector that reflects the textual context, 3) a pooling layer, which transforms an RNN vector sequence to a distributed representation vector, and 4) a linear layer with sigmoid activation, which projects the distributed representation vector to a scalar score value.

The BERT-based model performs ASAG by executing the following procedures: 1) Add a special [CLS] token to the beginning of each answer text. 2) Add an output layer (typified by the linear layer with sigmoid activation) over the output vector corresponding to the [CLS] token. 3) Fine-tune this BERT-based ASAG model using a training dataset that consists of scored short-answer texts.

This work was supported by JSPS KAKENHI Grant Numbers 19H05663, 20K20817, and 21H00898. We thank Yuki Doka and Yoshihiro Kato at the Benesse Educational Research & Development Institute for providing the data used in this study.

2161-377X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICALT58122.2023.00104

The training of those conventional DNN-based ASAG models is generally conducted independently for each short-answer question. Therefore, they cannot consider any shared factors that affect examinees' performance on multiple questions. The goal of this study is to examine the effectiveness of examinee-specific latent features, which are expected to reflect latent traits in examinees, as shared features across multiple questions.

III. Proposed model

We propose a new DNN-based ASAG model with an examinee-aware architecture consisting of neural layers shared across multiple questions that extracts examinee-specific latent features. Specifically, the proposed model is formulated as a multi-input and multi-output DNN model that receives an examinee's short-answer texts for multiple questions on a single test and simultaneously outputs multiple corresponding scores for those input texts. The proposed model consists of the following three components.

1) **Text encoder.** This component transforms the word sequence in a given short-answer text to a distributed representation, a fixed-length feature vector for the text, through a DNN architecture. Specifically, assume that there are Q short-answer questions in a target test and let an examinee's answer text for question q be $\boldsymbol{t}_q = \{\boldsymbol{w}_{ql} \mid l \in \{1, 2, \ldots, L_q\}\}$, where $\boldsymbol{w}_{ql}$ indicates the l-th word in the text and L_q represents the number of words in the text. Then, the text encoder transforms the examinee's short-answer texts $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_Q\}$ into distributed representation vectors $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_Q\}$, respectively, where $\boldsymbol{F}_q$ indicates a distributed representation vector for $\boldsymbol{t}_q$.
2) **Examinee-aware architecture.** This component is a shared architecture across multiple questions that extract examinee-specific latent features. This component first concatenates the distributed representations of answer texts for questions $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_Q\}$. Next, the fully connected layer with hyperbolic tangent activation is applied to transform the concatenated vector to a lower-dimensional hidden vector $\boldsymbol{E}$, which can be regarded as examinee-specific features. Specifically, $\boldsymbol{E} = \tanh(\boldsymbol{W}^e[\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_Q] + \boldsymbol{b}^e)$, where $\boldsymbol{W}^e$ is a weight matrix, $\boldsymbol{b}^e$ is a bias vector, and $[\cdot, \ldots, \cdot]$ indicates the concatenation operation of the vectors. Finally, the examinee-specific feature vector $\boldsymbol{E}$ is concatenated with each text-specific distributed representation vector $\boldsymbol{F}_q$, and the concatenated vector $[\boldsymbol{E}, \boldsymbol{F}_q]$ is used to predict question-specific scores in the next output layer. Note that the last concatenation operation is important for appropriately extracting examinee-specific features. If the examinee-aware architecture passes only vector $\boldsymbol{E}$ to the output layer, $\boldsymbol{E}$ must simultaneously store examinee-specific features and question-relevant features. Meanwhile, because concatenation of $\boldsymbol{E}$ and $\boldsymbol{F}_q$ allows the output layer to refer directly to question-specific features $\boldsymbol{F}_q$, $\boldsymbol{E}$ can aggregate focusing only on features shared across the questions.

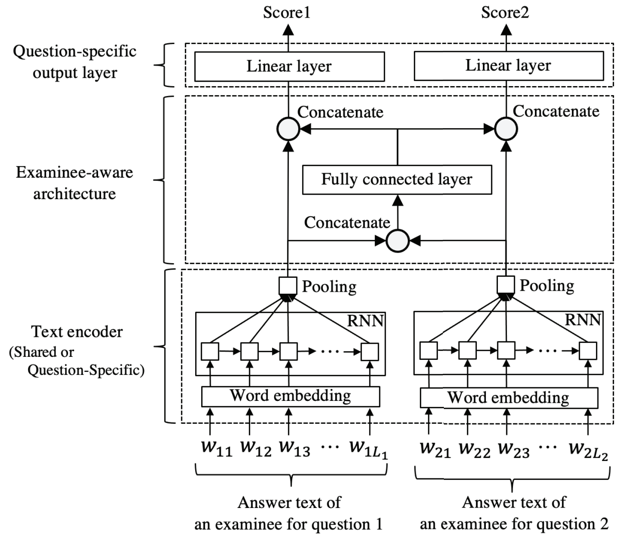
Score1
Score2
Question-specific output layer
Linear layer
Linear layer
Concatenate
Concatenate
Examinee-aware architecture
Fully connected layer
Concatenate
Pooling
Pooling
Text encoder (Shared or Question-Specific)
RNN
RNN
Word embedding
Word embedding
w_{11} w_{12} w_{13} $\cdots$ w_{1L_1}
w_{21} w_{22} w_{23} $\cdots$ w_{2L_2}
Answer text of an examinee for question 1
Answer text of an examinee for question 2

Fig. 1. Architecture of the proposed RNN-based model with two short-answer questions on a test.

3) **Question-specific output layer.** This component consists of question-specific linear layers in which a layer for question q projects the concatenated vector $[\boldsymbol{E}, \boldsymbol{F}_q]$ to a scalar score value. In this study, we apply a sigmoid activation, as in many conventional ASAG models. Specifically, the score of the answer text for q-th question is calculated as $\sigma\left(\boldsymbol{W}_q^o[\boldsymbol{E}, \boldsymbol{F}_q] + b_q^o\right)$, where $\boldsymbol{W}_q^o$ is a weight vector and b_q^o is a bias parameter in the output layer for Question q.

The proposed model can use various DNN architectures as the text encoder. To give an example, Figure 1 shows the architecture of the proposed model using the RNN-based model as the text encoder. A BERT-based model would also be possible by switching the text encoder layer to BERT.

The proposed model is trained by a backpropagation algorithm using the mean squared error loss function $\frac{1}{Q \cdot N} \sum_{q=1}^{Q} \sum_{n=1}^{N} (y_{nq} - \hat{y}_{nq})^2$, where N indicates the number of examinees in a training dataset, and y_{nq} and $\hat{y}_{nq}$ indicate the true and predicted scores for the n-th examinee's short answer to question q. Note that the true scores in the training dataset are normalized to a $[0, 1]$ scale during the training phase, because sigmoid activation is used in the output linear layers. During the prediction phase, predicted scores are rescaled to the original score range.

IV. Experiments

In this section, the effectiveness of the proposed model is demonstrated through experiments using actual data.

A. Actual data

There are several open datasets that have commonly been used for ASAG research. However, they cannot directly be

TABLE I
PERFORMANCE OF THE PROPOSED AND CONVENTIONAL MODELS.

models	Question 1	Question 2	Question 3	*Avg.*
Conv. RNN	0.547	0.774	0.665	0.662
Prop. RNN	**0.628**	**0.831**	**0.705**	**0.721**
Conv. BERT	0.858	0.891	0.795	0.848
Prop. BERT	**0.880**	**0.907**	**0.834**	**0.874**

used for our study because they contain no examinee identifiers and no information for identifying which questions were offered on a single test. Therefore, for this experiment, we used data from a Japanese reading comprehension test developed by Benesse Educational Research and Development Institute in Japan. This dataset comprises responses to three short-answer questions given by 511 examinees (Japanese university students). The average number of characters in the short-answer texts for each question was 27, 33, and 50. The scores were given by expert raters, based on a scoring rubric with four grades for question 1 and other rubrics with five grades for questions 2 and 3.

B. Scoring accuracy evaluation

Using actual data, we evaluated the scoring accuracy of the conventional RNN- and BERT-based ASAG models as well as the proposed models using those ASAG models as text encoders. We evaluated scoring accuracy by five-fold cross-validation, using 60% of the data as training data, 20% as development data for selecting optimal hyperparameters and epoch, and the remaining 20% as test data. The accuracy metric is the quadratic weighted Kappa (QWK), which is commonly used in ASAG studies [1]. For the RNN-based models, we used long short-term memory as the RNN. Furthermore, the proposed RNN-based model was designed to use a shared word-embedding layer across multiple questions. For the BERT-based models, we used a *base*-sized BERT model pretrained on Japanese texts[1]. For the proposed models, we determined examinee-specific feature dimensions through a grid search and selected 20 dimensions as the optimal value.

Table I shows the experimental results, where the bold text indicates the best performance among the models with the same base ASAG model. Comparing the RNN- and BERT-based models, the BERT-based models show substantially higher accuracy, which is consistent with the findings of other studies [8]. Furthermore, this result shows that the proposed models provide higher accuracy compared with the corresponding conventional models for all questions, demonstrating the effectiveness of the proposed examinee-aware architecture.

C. Analysis of examinee-specific features

As explained earlier, examinee-specific features are expected to reflect latent examinee traits as a shared feature across multiple questions. This subsection thus examines the relation between examinee-specific features and latent examinee traits. In this analysis, we estimated the latent traits according to item response theory (IRT), a test theory based on mathematical models that have been widely used for educational and psychological measurements.

TABLE II
RELATIONSHIP BETWEEN IRT-BASED LATENT TRAITS IN EXAMINEES AND 20-DIMENSIONAL EXAMINEE-SPECIFIC FEATURES.

Dim	Correlation	Dim	Correlation	Dim	Correlation
1	-0.44*	8	0.07	15	0.10*
2	-0.29*	9	0.06	16	-0.11*
3	0.00	10	0.04	17	-0.22*
4	-0.01	11	0.16*	18	-0.24*
5	0.03	12	0.13*	19	-0.25*
6	0.10*	13	-0.15*	20	0.17*
7	-0.27*	14	-0.31*		

Table II shows the correlation between the IRT-based latent traits and the 20-dimensional examinee-specific features. Here, the examinee-specific features were obtained from the proposed RNN-based model. In Table II,* indicates that the p-value given by the significance test of the correlation coefficient was less than 0.05. According to Table II, many dimensions of the examinee-specific features were significantly correlated with the IRT-based latent traits, suggesting that examinee-specific features reflect latent examinee traits, although some dimensions might also reflect other shared factors.

V. CONCLUSION

In this paper, we proposed new DNN-based ASAG models with an examinee-specific feature extraction architecture that we called examinee-aware architecture. Actual data experiments demonstrate that use of the examinee-aware architecture effectively improves scoring accuracy. We also demonstrated that the examinee-specific features extracted by the proposed models reflect latent examinee traits.

REFERENCES

[1] S. Bonthu, S. R. Sree, and M. K. Prasad, "Automated short answer grading using deep learning: A survey," in *Mach. Learn. Knowl. Extraction*, 2021, pp. 61–78.

[2] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. M. Lee, "Investigating neural architectures for short answer scoring," in *Proc. Workshop Innovative Use of NLP for Building Educational Applications*, 2017, pp. 159–168.

[3] C. Sung, T. I. Dhamecha, and N. Mukhi, "Improving short answer grading using transformer-based pre-training," in *Proc. Int. Conf. Artificial Intelligence in Education*, 2019, pp. 469–481.

[4] L. Camus and A. Filighera, "Investigating transformers for automatic short answer grading," in *Proc. Int. Conf. Artificial Intelligence in Education*, 2020, pp. 43–48.

[5] D. Gautam and V. Rus, "Using neural tensor networks for open ended short answer assessment," in *Proc. Int. Conf. Artificial Intelligence in Education*, 2020, pp. 191–203.

[6] L. A. Ha, V. Yaneva, P. Harik, R. Pandian, A. Morales, and B. Clauser, "Automated prediction of examinee proficiency from short-answer questions," in *Proc. Int. Conf. Computational Linguistics*, 2020, pp. 893–903.

[7] M. Uto and Y. Uchida, "Automated short-answer grading using deep neural networks and item response theory," in *Proc. Int. Conf. Artificial Intelligence in Education*, 2020, pp. 334–339.

[8] Z. Li, Y. Tomar, and R. J. Passonneau, "A semantic feature-wise transformation relation network for automatic short answer grading," in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2021, pp. 6030–6040.

[1] https://huggingface.co/cl-tohoku/bert-base-japanese-v2