July 2024

Automatic Short-Answer Grading with a Pseudo-Siamese Neural Network

Tzu-Lin Chang
National Pingtung University of Science and Technology, tlchang@mail.npust.edu.tw

Keng-Pei Lin
National Sun Yat-Sen University, kplin@mis.nsysu.edu.tw

Zong-Shun Chen
National Sun Yat-Sen University, m104020056@student.nsysu.edu.tw

Automatic Short-Answer Grading with a Pseudo-Siamese Neural Network

Short Paper

Tzu-Lin Chang
National Pingtung University of Science and Technology
Pingtung, Taiwan
tlchang@mail.npust.edu.tw

Keng-Pei Lin
National Sun Yat-Sen University
Kaohsiung, Taiwan
kplin@mis.nsysu.edu.tw

Zong-Shun Chen
National Sun Yat-Sen University
Kaohsiung, Taiwan
m104020056@student.nsysu.edu.tw

Abstract

This study aims to develop an automatic short-answer grading system. The focus is on achieving few-shot learning in short-answer grading, tackling the scarcity of labeled data. We utilize a two-head pseudo-Siamese neural network with an external knowledge model for transfer learning to enhance the system's ability to extract key information in the student answers for the automatic grading. The experiments are performed on real data, and the results show that the proposed system can effectively achieve the objective with a high accuracy.

Keywords: Automatic grading, natural language processing, digital learning

Introduction

A short-answer question is a type of assessment which requires examinees to write a brief and concise response, typically consisting of several sentences. This type of questions can assess students' understanding of specific concepts in a subject, evaluate their ability to articulate concepts, and encourage them to actively demonstrate their knowledge. Unlike multiple-choice questions, short-answer questions do not provide answer options, requiring students to generate their own responses, which can demonstrate their grasp of the subject matter more effectively.

However, unlike multiple-choice questions in exams that can be automatically scored by computers, short-answer questions in exams are typically graded by teachers. Evaluating students' short-answer exams poses a significant challenge for teachers, whether it's a small-scale classroom quiz, a final exam, or a large-scale entrance examination. Teachers are required to maintain consistent and objective grading standards while assessing responses from hundreds or thousands of students. The varied answering styles of students in short-answer questions, with each student having their own way of expression, make automated grading of short answers even more challenging.

This work aims to develop an automatic short-answer grading system. Although several previous research works had addressed the automatic grading of short-answer questions (Siddiqi, Harrison, & Siddiqi, 2010; Süzen, Gorban, Levesley, & Mirkes, 2020; Mohler, Bunescu, & Mihalcea, 2011; Sung et al., 2019, Sultan, Salazar, & Sumner, 2016), there are still challenges. One is the need for a large amount of labeled data to

train machine learning models. Because the questions and answers for each examination should strive to avoid similarities with previous exams, it is challenging to obtain a large number of similar samples for training purposes. The other is that natural language understanding remains a challenging research area, particularly in comprehending and evaluating complex perspectives and arguments in students' responses.

We adopt a two-head pseudo-Siamese neural network (Bromley, Guyon, LeCun, Säckinger, & Shah, 1994) to develop the short-answer grading system. The Siamese neural network is distinctive as it has two or more homogeneous inputs sharing the same network, and it can assess the similarity between inputs based on their differences in a feature space. Through the comparative approach, the Siamese neural network can effectively learn patterns from a small amount of training samples for generalizing to new samples.

The objective of this work is to develop an automatic short-answer grading system based on few-shot learning, capable of handling grading with a small number of training samples. Due to the difficulty of acquiring a complete knowledge structure with limited labeled data, we introduce external knowledge model in addition to the question-answer sets. Transfer learning will be applied to transfer the acquired knowledge to the task of automatic short-answer grading, enhancing the system's ability to extract key information and better handle short-answer grading in specific domains.

In summary, we propose an approach for learning an automatic short-answer grading system with a small question-answer dataset. We capitalize on a two-head pseudo-Siamese neural network to learn the grading of student answers. One head of the network is to learn features directly from the question-answer dataset, and the other head is a pretrained external knowledge model for transfer learning the knowledge obtained from textbook or Internet resources for enriching the features to train on a small question-answer dataset. The proposed automatic short answer grading system can be applied in educational, examination, and other testing scenarios, with the potential to reduce grading costs and enhance grading efficiency.

Literature Review

In the early stages of automatic short-answer grading, rule-based systems were commonly used (Wresch, 1993). These systems relied on predefined rules and heuristics to evaluate responses. However, they were limited in handling variations in language and lacked adaptability. With advancements in natural language processing (NLP), researchers began exploring techniques such as text analysis, syntactic and semantic parsing, and machine learning algorithms to automatically grade short answers (Siddiqi, Harrison, & Siddiqi, 2010; Süzen, Gorban, Levesley, & Mirkes, 2020; Mohler, Bunescu, & Mihalcea, 2011). NLP allows systems to understand the context, semantics, and structure of written responses. Supervised machine learning models are trained on labeled datasets of human-graded responses to learn patterns and associations between features and grades.

Feature extraction plays a crucial role in automatic short-answer grading (Burrows, Gurevych, & Stein, 2015). Researchers have explored various linguistic and contextual features, such as word frequency, grammatical structure, and coherence, to improve the accuracy of grading systems. Semantic analysis involves understanding the meaning and intent behind a response. Some approaches use semantic similarity measures or ontologies to compare student answers with model answers, allowing for a more nuanced evaluation.

Deep learning techniques, particularly recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have gained popularity in short-answer grading (Tulu, Ozkaya, & Orhan, 2021; Wu & Yeh, 2019). These models can capture intricate patterns in text data and automatically learn hierarchical representations.

The machine learning-based automatic short-answer grading approaches usually require a large amount of training data to achieve high performance, where it may be difficult to obtain the large-scale training data in real examination scenarios. In this work, we design a few-shot learning approach based on Siamese neural networks for the automatic short-answer grading.

Proposed Method

The system architecture of this research primarily consists of two models: the external knowledge model and the grading model. The grading model trains on existing question-answering dataset, and the external

knowledge model incorporates unstructured knowledge from Wikipedia to assist the grading model in better understanding and evaluating students' responses. Since in the process of short-answer grading, students' responses may involve information and knowledge beyond the scope of the course materials, making it challenging to evaluate their answers solely based on existing question-answer datasets. For this reason, we propose a strategy that combines external knowledge, allowing the grading model to integrate a diverse range of information from the Wikipedia to enhance the accuracy of grading.

The external knowledge model acquires one-dimensional convolutional features from external knowledge resources to enhance its understanding and analytical capabilities of input text. The hidden layers obtained from feature extraction will be utilized as a component in the grading model, which integrates a pre-trained external knowledge model and is fine-tuned with a question-answering dataset in conjunction with the training of the grading model.

The final short-answer grading model is a two-headed pseudo-Siamese neural network, composing of a deep head as a one-dimensional convolutional neural network, and a shallow head with fully connected layers. The deep head is to fine-tune the pre-trained external knowledge model, while the shallow head is for learning the features of the question-answering dataset. The fusion of these two heads will be fed to a fully connected layer to predict the grade of the answer.

To process the unstructured textual input, we adopt LASER sentence embedding (Artetxe & Schwenk, 2019) to generate vector representations for sentences. Each input sentence to LASER will be mapped to a 1024-dimensional vector. The sentences with similar meaning will be closer in the semantic vector space. Because the use of sentence embedding, the traditional text preprocessing steps such as stop word removal or part-of-speech tagging, etc., are not required. The original sentence can be inputted directly.

The architecture of the proposed approach is shown in Figure 1. Each training instance consists of a pair of reference answer and student answer and will be converted to sentence embeddings. The difference of the two embeddings will be used as the input to the pseudo-Siamese neural network.

We use Wikipedia to obtain relevant information in a professional field to build a data set for training an external knowledge model for transfer learning to the short-answer grading model. Part of the content of each entry in Wikipedia will be paired with different parts of the same page, regarded as a highly similar positive sample training instance, and is marked as 1. In addition, a piece of text is randomly selected from an irrelevant entry will be paired as a piece of negative training instance and is marked as 0. Considering that many responses may not be related to the correct answer, this study uses many random irrelevant texts paired with the original content to create many negative samples. These pairs of articles are like the pairs of reference answers and student answers for training the short-answer model. All the articles will be first transformed to vectors by LASER sentence embedding. Then the differences of all the articles pairs are utilized as the inputs for training a one-dimensional convolutional network, which consists of a one-dimensional convolutional layer followed by a max pooling layer and two fully connected layers, and then a fully connected output layer. By training the dataset with this one-dimensional convolutional neural network, we obtain the convolutional features for using as a pre-trained external knowledge model. Except the output layer, the four hidden layers of the one-dimensional convolutional network will replace one head of the Siamese neural network for short-answer grading, as shown in the left head of the pseudo-Siamese neural network in Figure 1.

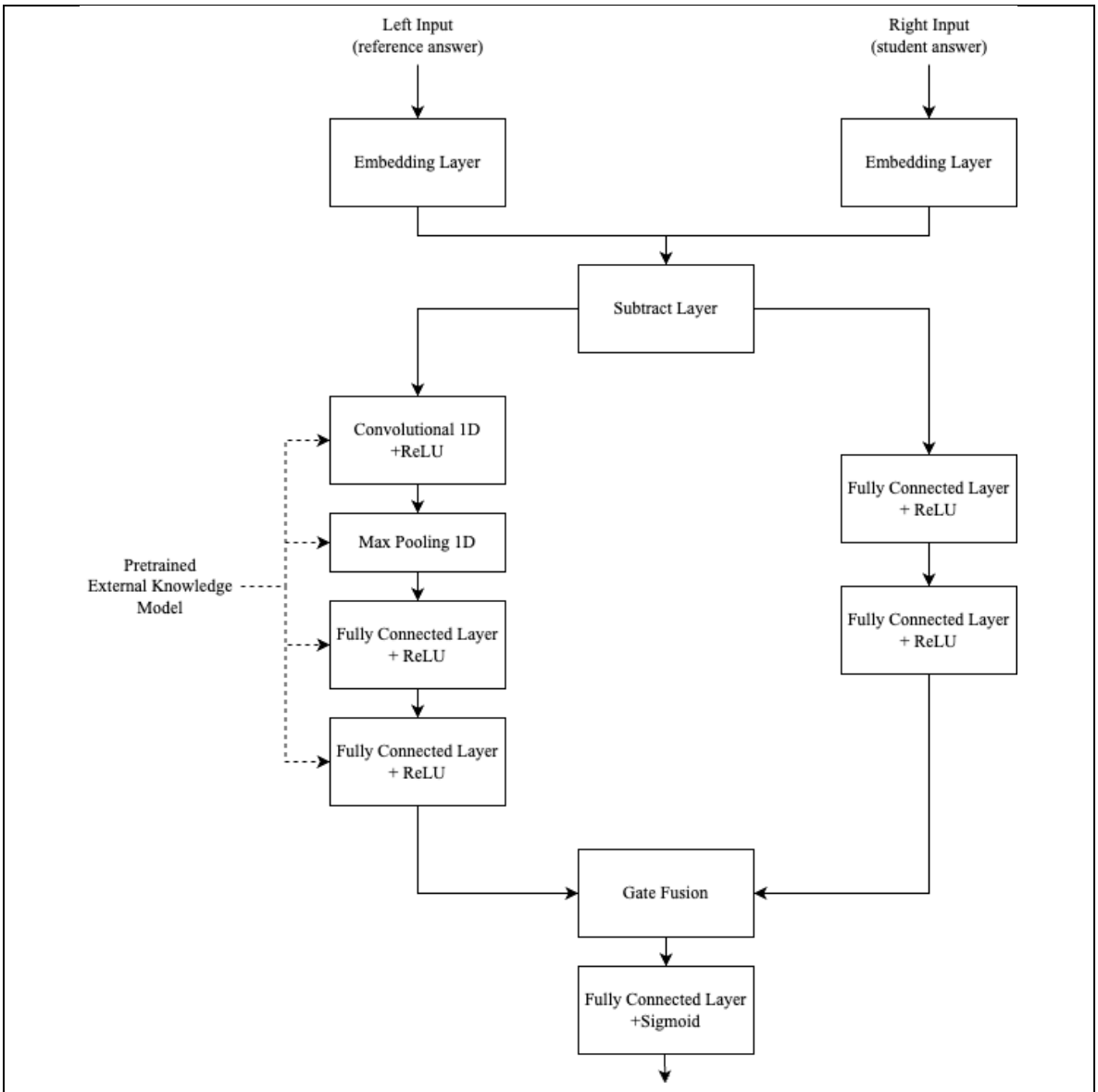

Figure 1. Proposed Network Architecture

Experiments

To evaluate the proposed automatic short-answer grading method, we experiment on the dataset from the data structures course of the University of North Texas (Mohler, Bunescu, & Mihalcea, 2011). This dataset comprises 81 questions, totaling 2,273 student responses collected from 31 students in examinations. Each question has a corresponding reference answer, and all responses are written in English. The grades range from 0 to 5, with each student's answer independently assessed by two human judges. The objective of the experiment is to predict the average grade given by the two human judges. The dataset is divided to the training set with 70% of the whole dataset, the validation set with 10% of the whole dataset, and the testing set with 20% of the whole dataset.

The proposed method is implemented in Python with the deep learning framework PyTorch and LASER library. The details of the network structure are shown in Table 1. During the training process, the NAdam

optimizer (Dozat, 2015) was employed with an initial learning rate of 0.001. The hyperparameters beta_1 and beta_2 were set to 0.9 and 0.999, respectively, and epsilon was set to 1e-7. A batch size of 512 was utilized, and dropout was set to 0.5. Additionally, early stopping technique was implemented to determine the optimal training epoch and mitigate the risk of overfitting.

Table 1. Details of the Network Structure

Part	Left Layers	Output Dimension	Number of Parameters
Input	Embedding	1024	-
	Subtract	1024	-
Left Head	Conv1D+ReLU	1022, 64	256
	Max Pooling 1D	511, 64	-
	Flatten	32704	-
	Fully Connected + ReLU	256	8,372,480
	Fully Connected + ReLU	128	32,896
Right Head	Flatten	1024	-
	Fully Connected + ReLU	256	262,400
	Fully Connected + ReLU	128	32,896
Output	Concatenate	256	-
	Fully Connected + Sigmoid	1	257
	Gate Fusion	128	-
	Fully Connected + Sigmoid	1	129

We compare the performance with the short-answer grading approaches proposed in the works of Sultan, Salazar, & Sumner (2016), which utilized the measures of text similarity between reference answers and student responses to trains a ridge regression model, and Mohler, Bunescu, & Mihalcea (2011), which trained a support vector regression model on the features extracted by measuring the lexical semantic similarity and dependency graph alignment. The results with Pearson'r and root mean square error (RMSE)

are shown in Table 2. The results show that the proposed method outperforms the previous works in both Pearson'r and RMSE.

Table 2. Performance Comparison

Method	Pearson's r ↑	RMSE ↓
Mohler et al. (2011)	0.518	0.978
Sultan et al. (2016)	0.630	0.850
Proposed method	**0.774**	**0.705**

Conclusion

We propose an automated short-answer grading system, emphasizing the attainment of few-shot learning to address limited labeled data. By employing a two-head pseudo-Siamese neural network, augmented with an external knowledge model for transfer learning, we demonstrate the system effectiveness by comparing the results with previous works.

Since the proposed short-answer grading model is designed to learn from a small question-answer set, it relies on transfer learning from an external knowledge model. Therefore, a large number of articles related to the subject are required. If there are insufficient resources for training the external knowledge model, the proposed model will not be able to learn sufficiently rich features from the small question-answer set, leading to poorer performance. The pre-trained language model such as BERT (Devlin, Chang, Lee, & Toutanova, 2019) may benefit the training of the external knowledge model. In future works, we plan to exploit on fine-tuning the BERT language model for generating the external knowledge model to be combined with the Siamese neural networks. Furthermore, we want to extend the proposed method to cross-language short-answer grading by utilizing the language-agnostic sentence embeddings.

References

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597–610. **https://aclanthology.org/Q19-1038**

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1994). Signature verification using a "Siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6, 737–744. **https://proceedings.neurips.cc/paper/1993/file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf**

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25, 60–117. **https://doi.org/10.1007/s40593-014-0026-8**

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, Volume 1 (Long and Short Papers), 4171-4186.* **https://aclanthology.org/N19-1423/**

Dozat, T. (2015). Incorporating Nesterov momentum into Adam. Retrieved 2024-05-15. **https://cs229.stanford.edu/proj2015/054_report.pdf**

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 752-762. **https://aclanthology.org/P11-1076/**

Siddiqi, R., Harrison, C. J., & Siddiqi, R. (2010). Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies*, 3(3), 237-249. **https://doi.org/10.1109/TLT.2010.4**

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, 169, 726-743. **https://doi.org/10.1016/j.procs.2020.02.171**

Sultan, M. A., Salazar, C., & Sumner, T. (2016). Fast and easy short answer grading with high accuracy. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1070-1075. **https://aclanthology.org/N16-1123**

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019). Pre-training BERT on domain resources for short answer grading. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6071-6075. **https://aclanthology.org/D19-1628**

Tulu, C. N., Ozkaya, O., & Orhan, U. (2021). Automatic short answer grading with SemSpace sense vectors and MaLSTM. *IEEE Access*, 9, 19270-19280. **https://doi.org/10.1109/ACCESS.2021.3054346**

Wresch, W. (1993). The imminence of grading essays by computer-25 years later. *Computers and Composition*, 10(2) 45-58. **https://doi.org/10.1016/S8755-4615(05)80058-1**

Wu, S.-H., & Yeh, C.-Y. (2019). A short answer grading system in Chinese by CNN. *Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*. **https://doi.org/10.1109/ICAwST.2019.8923528**