Opening the Black Box: Unraveling the Classroom Dialogue Analysis (Student Abstract)

Deliang Wang

Faculty of Education, The University of Hong Kong, Hong Kong, China
wdeliang@connect.hku.hk

Abstract

This paper explores proposing interpreting methods from explainable artificial intelligence to address the interpretability issues in deep learning-based models for classroom dialogue. Specifically, we developed a Bert-based model to automatically detect student talk moves within classroom dialogues, utilizing the *TalkMoves* dataset. Subsequently, we proposed three generic interpreting methods, namely saliency, input*gradient, and integrated gradient, to explain the predictions of classroom dialogue models by computing input relevance (i.e., contribution). The experimental results show that the three interpreting methods can effectively unravel the classroom dialogue analysis, thereby potentially fostering teachers' trust.

Introduction

Classroom dialogue plays a crucial role in shaping the quality of teaching and learning, as classroom utterances contain valuable information about learning. Effectively and promptly capturing this key information enables teachers to provide adaptive teaching. To achieve this goal, researchers have developed many teacher professional development programs aimed at equipping teachers with the necessary skills to leverage classroom dialogue and enhance their teaching practices. Empirical studies have extensively demonstrated the efficacy of such programs. However, these programs face a significant limitation: teachers are usually required to manually analyze classroom recordings to improve their dialogic skills. This manual analysis poses overwhelming challenges as it is time-consuming and labor-intensive for teachers to extract and analyze key segments from the extensive corpus of classroom dialogue. Consequently, researchers have turned to artificial intelligence (AI) as a promising avenue for automated, immediate, and accurate analysis in this domain.

Traditional machine learning techniques (e.g., random forest and decision tree) have predominantly been employed to investigate classroom dialogue. While these models provide a basic analysis, their performance remains limited. Subsequently, researchers have utilized deep learning techniques (e.g., RNN and CNN) to examine classroom dialogue. It is widely acknowledged that deep learning-based models offer more accurate analysis. However, these models often exhibit complex structures and operate as black boxes, lacking clear explanations for the analysis of teachers' or students' utterances. This lack of transparency and interpretability can lead to teachers distrusting and even refusing to use such models. Furthermore, incorrect analysis of classroom dialogue without explanations may result in teachers making inappropriate adjustments to their teaching strategies, thereby undermining teaching quality. In the field of education, which prioritizes the accountability and trustworthiness of AI-powered tools, it is significant to address the interpretability issue. Recently, researchers have also explored using the latest large language model (e.g., ChatGPT) to analyze classroom dialogue. While ChatGPT offers a certain degree of explainability in its answers, there is still significant room for improving its performance in investigating classroom dialogue, particularly when compared to other deep learning-based models (Wang et al. 2023).

The field of Explainable AI (xAI) has presented numerous interpreting methods to unravel the decision-making process of intricate models. Nevertheless, limited attention has been given to the interpretability challenges within deep learning-based classroom dialogue models. Therefore, this study explores using three model-agnostic (i.e., generic) interpreting methods to unveil the analysis of classroom dialogue derived from deep learning models. The primary objective is to provide teachers with explanations regarding the analysis of students' utterances and foster their trust.

Method

Talk moves are specific dialogic acts that elicit responses and have demonstrated empirical efficacy in enhancing teaching and learning. In light of this, we have selected the *TalkMoves* dataset (Suresh et al. 2022) for constructing and elucidating deep learning-based models for analyzing student utterances within classroom dialogue. This dataset comprises 567 transcripts from K-12 mathematics lessons, with a total of 174,186 teacher utterances and 59,874 student utterances. Each student utterance is annotated with a label (i.e., five talk moves in total). We randomly allocated 90% of the data for training, while the remaining 10% was reserved for testing.

Given that our objective is not to develop a model with state-of-the-art performance, we opted to construct the classroom dialogue model for student talk move analysis us-

ing the most widely utilized *BertForSequenceClassification* model, based on existing literature. Considering the significance of social interaction in talk move analysis, we set the model input as the utterance pairs (i.e., a student utterance concatenated with its previous sequence). During the training process, we set the epoch, optimizer, batch size, and learning rate to 10, Adam, 32, and 2e-5. Finally, the student model achieved an F1 score of 0.67 and an accuracy of 0.78. Since our primary focus lies in model interpretability, we did not conduct cross-validation or performance optimization procedures.

Considering the diverse structures in classroom dialogue models, we employed three generic xAI methods: saliency, input*gradient, and integrated gradient. These methods enable interpretation of predictions by quantifying the relevance of each word or token. Formally, let x be an utterance with its embedding denoted as e ($e \in R^m$). The utterance x consists of n tokens, with each token's embedding represented as e_i, where i denotes the token position. The Bert model f predicts the probability $f_c(e)$ that the utterance belongs to talk move c. To compute the word relevance, the input*gradient method (Li et al. 2016) decomposes the nonlinear prediction $f_c(e)$ into a linear sum of token contributions, as Equation 1 shows. The dot product of i-th token's embedding e_i and its derivative $\frac{\partial f_c(e)}{\partial e_i}$ is treated as the token's relevance, as Equation 2 shows, where j represents the j-th dimension in e_i and m is the total number of dimensions. The saliency method (Simonyan, Vedaldi, and Zisserman 2013) takes the derivative $\frac{\partial f_c(e)}{\partial e_i}$ as a measure of the token's importance to the prediction, as Equation 3 shows.

$$f_c(e) \approx \sum_{i=1}^{n} \frac{\partial f_c(e)}{\partial e_i} \cdot e_i + b \tag{1}$$

$$R_{input}(e_i) \approx \sum_{j=1}^{m} \frac{\partial f_c(e)}{\partial e_{ij}} e_{ij} \tag{2}$$

$$R_{saliency}(e_i) \approx \sum_{j=1}^{m} \frac{\partial f_c(e)}{\partial e_{ij}} \tag{3}$$

To explain the prediction $f_c(e)$, the integrated gradient method chooses a baseline sample x' with an embedding of e' and a predicted probability $f_c(e')$. By comparing $f_c(e)$ and $f_c(e')$, the method attributes the difference to disparities in their token embeddings, as shown in Equation 4, where $R_{ig}(e_i)$ represents the relevance of token e_i to its prediction. The integrated gradient method considers a linear path from the baseline embedding e' to the input embedding e and computes the gradients at every point along the path (Sundararajan, Taly, and Yan 2017). The accumulation of these gradients yields $R_{ig}(e_i)$, as Equation 5 shows. Here, we set the baseline sample as a sample with all-zero tokens. By obtaining the token relevance, we can present to teachers how the dialogue analysis is made.

$$\sum_{i=1}^{n} R_{ig}(e_i) = f_c(e) - f_c(e') \tag{4}$$

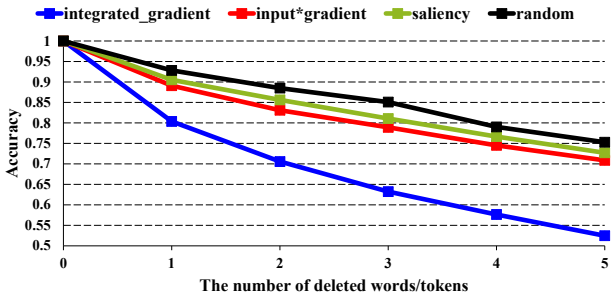

Figure 1: The accuracy changes of deleting tokens from correctly-predicted utterances based on their relevance

$$R_{ig}(e_i) \approx \sum_{j=1}^{m} \left(e_{ij} - e'_{ij}\right) \times \int_{\alpha=0}^{1} \frac{\partial f\left(e' + \alpha \times (e - e')\right)}{\partial e_{ij}} d\alpha \tag{5}$$

Results

To determine whether the identified important tokens were indeed crucial for predicting the talk move of student utterances, we conducted a validation experiment. Specifically, we selected utterances that had been correctly predicted with 100% accuracy, and then removed important tokens (i.e., set them to zeros) in a decreasing order of their relevance, one by one, until five tokens had been removed. By examining the changes in accuracy, we were able to confirm the effectiveness of token relevance. Given that deleting tokens from a short utterance could significantly impact prediction accuracy, we set a minimum length of 10 and also included a random deletion for comparison.

Figure 1 shows the prediction accuracy change among the initially correctly-predicted utterances. Notably, removing tokens based on their relevance computed from the three interpreting method results in a more significant decline in accuracy for the student model compared to random deletion. This validates the effectiveness of the interpreting results. The integrated gradient method, as the most effective one, can be used in providing more accurate explanations.

Acknowledgements

This work was supported by the Natural Science Foundation of China (No. 62277045), Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221), and Innovation and Technology Commission of the Government of the HKSAR (Grant No.: ITB/F-BL/7026/20/P)..

References

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2016. Visualizing and Understanding Neural Models in NLP. In *International Conference of the North American Chapter of the Association for Computational Linguistics*, 681–691.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image

classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, 3319–3328. PMLR.

Suresh, A.; Jacobs, J.; Harty, C.; Perkoff, M.; Martin, J. H.; and Sumner, T. 2022. The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4654–4662.

Wang, D.; Shan, D.; Zheng, Y.; Guo, K.; Chen, G.; and Lu, Y. 2023. Can ChatGPT Detect Student Talk Moves in Classroom Discourse? A Preliminary Comparison with Bert. In *Proceedings of the 16th International Conference on Educational Data Mining*, 515–519. International Educational Data Mining Society.