Automatic Dialogic Instruction Detection for K-12 Online One-on-One Classes

Shiting Xu, Wenbiao Ding, and Zitao Liu(✉)

TAL Education Group, Beijing, China
{xushiting,dingwenbiao,liuzitao}@100tal.com

Abstract. Online one-on-one class is created for highly interactive and immersive learning experience. It demands a large number of qualified online instructors. In this work, we develop six dialogic instructions and help teachers achieve the benefits of one-on-one learning paradigm. Moreover, we utilize neural language models, i.e., long short-term memory (LSTM), to detect above six instructions automatically. Experiments demonstrate that the LSTM approach achieves AUC scores from 0.840 to 0.979 among all six types of instructions on our real-world educational dataset.

Keywords: Dialogic instruction · One-on-one class · K-12 education · Online education

1 Introduction

With the recent development of technology such as digital video processing and live streaming, various forms of online classes emerge [4]. Because of the better accessibility and live learning experience, one-on-one class stands out where students are able to not only study materials at their only own pace, but have opportunities to frequently interact with their teachers facially and vocally [3,13,15]. Online one-on-one class has demonstrated its personalized education experience as supplements to the traditional training from public schools [14].

In spite of the above benefits, online one-on-one classes pose numerous challenges on instructors. On one hand, the instructor qualifications are significantly different from those in public schools. Public school teachers focus on making sure that the majority students are on track and pass their qualification examinations. While one-on-one instructors need to pay detailed attentions to every single student and adjust their teaching paces, styles, or even contents accordingly. Furthermore, students enroll in one-on-one courses for high-frequency interactions. This requires the teachers to encourage and lead students' active participations. On the other hand, a large portion of one-on-one participants are academically low-ranking K-12 students. Most of them are eager to study but don't know how to learn. The one-on-one instructors are responsible to help them build effective study habits. Therefore, in order to scale the qualified supply of one-on-one instructors and provide more effective and personalized education to the general

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 340–345, 2020.
https://doi.org/10.1007/978-3-030-52240-7_62

K-12 students, we develop six in-class dialogic instructions for one-on-one class teachers. Moreover, we build an end-to-end system to automatically detect and analyze the proposed pedagogical instructions.

2 Related Work

Many existing methods have been developed to analyze classroom dialogic instructions. Wang et al. identify teacher lecturing, class discussion and student group work in the traditional classroom by asking teachers to wear the LENA system [8] during the class [22]. Donnelly et al. identify occurrences of some key instructional segments, such as Question & Answer, Supervised Seatwork, etc., by using Naive Bayes models [5]. Owens et al. develop Decibel Analysis for Research in Teaching, i.e., DART, to analyzes the volume and variance of classroom recordings to predict the quantity of time spend on single voice (e.g., lecture), multiple voice (e.g., pair discussion), and no voice (e.g., clicker question thinking) activities [17].

Our work is distinguished from existing research studies because (1) we focus on the K-12 online one-on-one domain and propose six pedagogical instructions explicitly designed for it; (2) our dialogic instruction detection approach is an end-to-end solution that doesn't require any human intervention or any additional recording device.

3 Our Approach

3.1 Dialogic Instructions

By analyzing thousands of online one-on-one class videos and surveying hundreds of instructors, students, parents and educators, we categorize six dialogic instructions for K-12 online one-on-one classes as follows:

- *greeting:* Greeting instructions help teachers manage their teaching procedures before the class, such as greeting students, testing teaching equipments. Examples: "How are you doing?", "Can you hear me?", etc.
- *guidance:* Guidance instructions ask teachers to interact with students when lecturing on a particular knowledge point or a factual answer. Examples: "Do you know the reason?", "Let's see how we can get there?", etc.
- *note-taking:* Note-taking instructions require teachers to help students learn how to take notes and assist them to build effective learning habits. Examples: "Highlight this paragraph.", "Please copy this part", etc.
- *commending:* Commending instructions ask teachers to encourage students and build their confidence. Examples: "Good job.", "Well done.", etc.
- *repeating:* Repeating instructions remind teachers to let students retell the content by themselves, which enhances their understandings. Examples: "Could you please explain that to me?", "Can you rephrase that?", etc.
- *summarization:* Summarization instructions ask teachers to summarize teaching contents and materials at the end of the each class and conclude the main takeaways. Examples: "Let's review the key points", "Let's wrap up.", etc.

3.2 The Dialogic Instruction Detection Approach

The end-to-end dialogic detection pipeline takes class recordings as input and outputs spoken sentences of the above six types of dialogic instructions. The entire workflow is illustrated in Fig. 1, which consists of two key components: *Audio Processing* and *Language Modeling*.

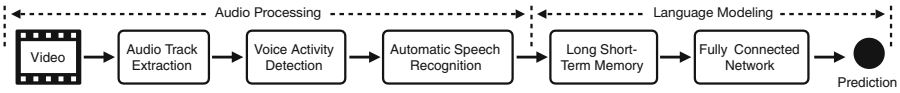

Fig. 1. The workflow of the end-to-end dialogic instruction detection approach.

Audio Processing. Audio processing involves three key steps: (1) extracting audio tracks from video recordings; (2) cutting audio tracks into short-span segments and removing noises and silence segments by a voice activity detection (VAD) algorithm; and (3) transcribing each audio segment by using an automatic speech recognition (ASR) algorithm. Please note that since both students' and teachers' videos are recorded separately, voice overlaps don't exist in the video recordings. This avoids the unsolved challenge of speaker diarization [1,21].

Language Modeling. We conduct language modeling on the transcriptions from the audio processing module. For each word, we first fetch its low dimensional embeddings from a pre-trained word2vec model. After that, we build neural classifiers for each type of dialogic instructions defined in Sect. 3.1. In this work, we use the long short-term memory (LSTM) as our language modeling networks [9,10]. The LSTM models take a sentence as input and sequentially update the hidden state representation of each word by using a well designed memory cell, which is able to capture the long range dependencies within each sentence. The details of LSTM can be found in [9,10]. LSTM model have been successful in language modeling tasks such as text classification [12,24], machine translation [23], etc. Finally, we build a two-layer fully-connected position-wise feed forward network on the last hidden representation of LSTM to conduct the final predictions.

4 Experiments

In this work, we collect 2940 sentences for each type of dialogic instruction by manually annotating class recordings from a third-party online one-on-one learning platform[1]. Each sentence is associated with a binary label, indicating whether the sentence belongs to a dialogic instruction. We use 2352 sentences for

[1] https://www.xes1v1.com/.

training and the rest for validation and testing. Similar to Blanchard et al. [2], we find that publicly available AI engines may yield inferior performance in the noisy and dynamic classroom environments. Therefore, we train our own VAD [20], ASR [25] and word2vec [16] models on the classroom specific datasets.

We compare the LSTM language modeling network with several widely used baselines: logistic regression [11], i.e, *LR*, support vector machine [18], i.e., *SVM*, and gradient boosting decision trees [7], i.e., *GBDT*. Similar to Tang et al. [19], we use area under curve (AUC) score to evaluate the model performance [6].

4.1 Model Performance

Figure 2 shows that our LSTM approach outperforms all other methods on all six types of dialogic instruction detection tasks. Specifically, from Fig. 2, we find that simple instructions are relatively fixed and have little variants, such as "greeting" and "summarization". All the approaches have comparable performance. While for complex instructions with many language variations such as "note-taking", "commending" and "repeating", *LSTM* significantly outperforms other baselines by large margins. We believe this is because the sequential neural networks are able to capture the long contextual language dependence within the sentence, which is very important when dealing with colloquial conversations.

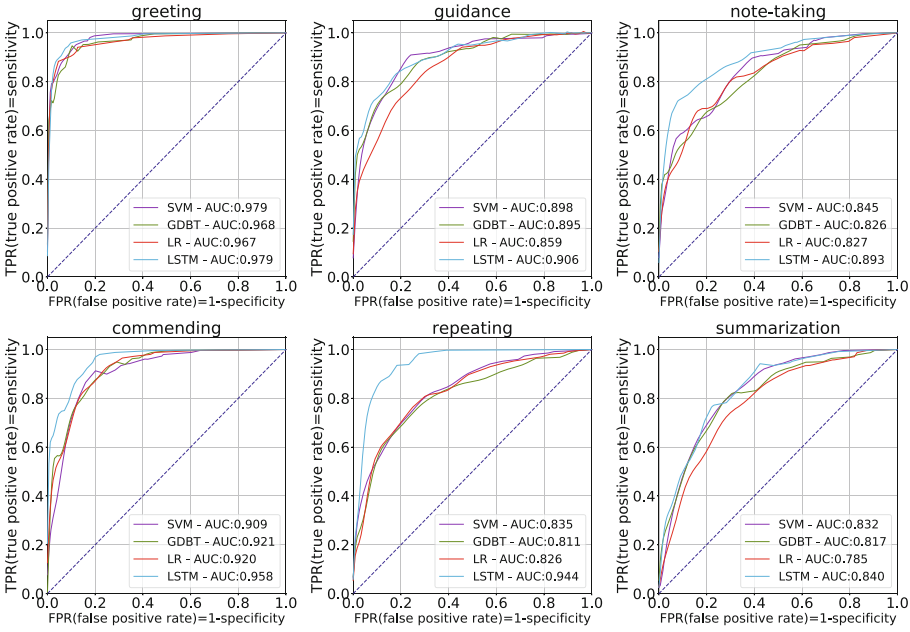

Fig. 2. ROC curves for detection performance of six dialogic instructions.

5 Conclusion

In this work, we propose six dialogic instructions and build an end-to-end solution for online one-on-one instructors. Experiments on a real educational dataset show that our LSTM based approach outperforms other baselines in the proposed six dialogic instructions.

References

1. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: a review of recent research. IEEE Trans. Audio Speech Lang. Process. **20**(2), 356–370 (2012)
2. Blanchard, N., et al.: A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 23–33. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19773-9_3
3. Chen, J., Li, H., Wang, W., Ding, W., Huang, G.Y., Liu, Z.: A multimodal alerting system for online class quality assurance. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11626, pp. 381–385. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23207-8_70
4. China Education Resources: The largest education system in the world is going online (2012). http://www.chinaeducationresources.com/s/OurMarket.asp. Accessed 5 Feb 2019
5. Donnelly, P.J., et al.: Automatic teacher modeling from live classroom audio. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, pp. 45–53. ACM (2016)
6. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
7. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
8. Ganek, H., Eriks-Brophy, A.: The language environment analysis (lena) system: a literature review. In: Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16 November 2016, pp. 24–32. No. 130, Linköping University Electronic Press (2016)
9. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2016)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
11. Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, vol. 398. Wiley, Hoboken (2013)
12. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI Conference on Artificial Intelligence (2015)
13. Li, H., et al.: Multimodal learning for classroom activity detection. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9234–9238. IEEE (2020)
14. Liang, J.K., et al.: A few design perspectives on one-on-one digital classroom environment. J. Comput. Assist. Learn. **21**(3), 181–189 (2005)

15. Liu, Z., et al.: Dolphin: a spoken language proficiency assessment system for elementary education. In: Proceedings of the Web Conference 2020, pp. 2641–2647. ACM (2020)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
17. Owens, M.T., et al.: Classroom sound can be used to classify teaching practices in college science courses. Proc. Natl. Acad. Sci. **114**(12), 3085–3090 (2017)
18. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
19. Tang, C., Ouyang, Y., Rong, W., Zhang, J., Xiong, Z.: Time series model for predicting dropout in massive open online courses. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 353–357. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93846-2_66
20. Tashev, I., Mirsamadi, S.: DNN-based causal voice activity detector. In: Information Theory and Applications Workshop (2016)
21. Tranter, S.E., Reynolds, D.A.: An overview of automatic speaker diarization systems. IEEE Trans. Audio Speech Lang. Process. **14**(5), 1557–1565 (2006)
22. Wang, Z., Pan, X., Miller, K.F., Cortina, K.S.: Automatic classification of activities in classroom discourse. Comput. Educ. **78**, 115–123 (2014)
23. Wu, Y., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
24. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)
25. Zhang, S., Lei, M., Yan, Z., Dai, L.: Deep-FSMN for large vocabulary continuous speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5869–5873. IEEE (2018)