Multi-Task Learning based Online Dialogic Instruction Detection with Pre-trained Language Models

Yang Hao[1], Hang Li[1], Wenbiao Ding[1], Zhongqin Wu[1], Jiliang Tang[2], Rose Luckin[3], Zitao Liu[1]⋆

[1] TAL Education Group, Beijing, China
{haoyang2,lihang4,dingwenbiao,wuzhongqin,liuzitao}@tal.com
[2] Data Science and Engineering Lab, Michigan State University, USA
tangjili@msu.edu
[3] UCL Knowledge Lab, London, UK
r.luckin@ucl.ac.uk

Abstract. In this work, we study computational approaches to detect online dialogic instructions, which are widely used to help students understand learning materials, and build effective study habits. This task is rather challenging due to the widely-varying quality and pedagogical styles of dialogic instructions. To address these challenges, we utilize pre-trained language models, and propose a multi-task paradigm which enhances the ability to distinguish instances of different classes by enlarging the margin between categories via contrastive loss. Furthermore, we design a strategy to fully exploit the misclassified examples during the training stage. Extensive experiments on a real-world online educational data set demonstrate that our approach achieves superior performance than other baselines. To encourage reproducible results, we make our code online available at https://github.com/AIED2021/multitask-dialogic-instruction.

Keywords: Dialogic instruction · Multi-task learning · Pre-trained language model · Hard example mining.

1 Introduction

Teaching online classes is a very challenging task for the well-trained offline classroom instructors. When sitting in front of a camera or a laptop, traditional classroom instructors are lack of effective pedagogical instructions to ensure the overall quality of their online classes. In this paper, we develop a set of dialogic instructions for online classes aiming to encourage talks and discourses between teachers and students, not just teacher-presentation [9, 14, 10, 6]. Furthermore, we study computational approaches to automatically detect these dialogic instructions from online class videos, which may provide timely feedback to teachers and help them improve their online teaching skills.

⋆ Corresponding Author: Zitao Liu

However, automatic dialogic instruction detection poses numerous challenges in real-life teaching scenarios. First, online teaching is not a standardized procedure. Even for the same learning content, different instructors may teach it in various ways according to their own pedagogical styles. Furthermore, the quality of dialogic instructions varies a lot from junior to senior instructors. The difference between effective and incompetent dialogic instructions is very subtle. The second challenge is that the model has to be robust enough to errors from automatic speech recognition (ASR) transcriptions. The publicly available ASR service may yield very high transcription errors and inferior performance in the noisy and dynamic classroom environments [3].

To address the above challenges, in this study, we propose an end-to-end multi-task framework for automatic dialogic instruction detection from online videos. Specifically, we (1) propose a contrastive loss based multi-task framework to distinguish instances by enlarging the distances between instances of different categories [12, 18]; (2) utilize the pre-trained neural language model to robustly handle errors from ASR transcriptions without manual annotation efforts [5, 15]; and (3) propose a strategy to select and exploit hard instances in the training process to achieve higher performance [21, 18].

2 The Dialogic Instruction Detection Framework

In this work, we aim to capture the following eight types of well-studied dialogic instructions that (1) motivate students and make them feel easy about the class: *greeting* [7, 16] and *commending* [10, 6], (2) help students understand learning materials and retain them: *guidance* [25], *example-giving* [19], *repeating* [2], and *reviewing* [1], and (3) build effective learning habits: *note-taking* [9, 14] and *summarization* [17].

Our multi-task dialogic instruction detection framework has three key components: (1) a pre-trained language model, which serves as the base model in the classification task; (2) a multi-task learning module, which distinguishes effective instructions from similar but ineffective ones by pushing instances from different categories apart; and (3) a hard example mining strategy, which establishes a hard example set to select instances when constructing input pairs.

Pre-trained Language Model: To extract contextual information, in this study we utilize the Transformer-based pre-trained language models as our base model in our detection framework. To perform the instruction detection task on a sentence, similar to [5, 15], we first add a special token $[CLS]$ in front of the sentence. After that, embeddings of each token in the sentence are fed into multiple Transformer encoders sequentially. Finally the hidden state of the special token $[CLS]$ from the last layer of Transformer encoders is obtained as the representation of the sentence.

Multi-task Learning Module: The multi-task learning framework consists of two sub-tasks: (1) a multi-class classification task to decide which category a dialogic instruction belongs to, where the cross-entropy loss is used; and (2) an additional task with an objective to enlarge the distances between pairs of instructions from different categories by using contrastive loss. The total loss is a combination of the two parts above defined as follows:

$$L = \gamma \cdot \underbrace{\sum_{i=1}^{b} \sum_{c \in \mathbf{C}} -y_i^c \cdot \log(\hat{y}_i)}_{\text{cross-entropy loss}} + (1-\gamma) \cdot \underbrace{\sum_{i=1}^{b} \left(\max\left\{0, M - \|\mathcal{F}_\Theta(\mathbf{x}_i) - \mathcal{F}_\Theta(\mathbf{x}_j^{\tilde{c}})\|_2\right\}\right)^2}_{\text{contrastive loss}}$$

where $\mathbf{x}_i$ denotes the raw feature of the ith instance and y_i^c represents the indicator variable that is equal to 1 if and only if the ith instance belongs to category c. $\hat{y}_i$ is the predicted label and b is the batch size. $\mathcal{F}_\Theta(\cdot)$ denotes the pre-trained language model, which extracts representation of an input instance. γ and M are hyper-parameters. $\mathbf{x}_j^{\tilde{c}}$ denotes an arbitrary instance (indexed by j) that comes from a different category of $\mathbf{x}_i$.

Hard Example Mining Strategy: Many instances that can be classified correctly by the model contribute little to the contrastive loss [18, 21]. That is to say, a randomly selected instance $\mathbf{x}_j^{\tilde{c}}$ probably has been far away from an instance $\mathbf{x}_i$ after epochs of training. Therefore, instead of generating pairs by random sampling, we focus on hard examples, i.e., instances that are misclassified into a wrong category. Hence, the hard example set $\mathbf{H}$ is discovered by: $\mathbf{H} = \{\mathbf{x}_j | \arg\max y_j \neq \arg\max \hat{y}_j, j = 1, \cdots, b\}$. Pairs of training inputs are selected by first randomly choosing an instance $\mathbf{x}_i$ from the entire training set $\mathbf{X}$, and then randomly choosing an $\mathbf{x}_j^{\tilde{c}}$ from the hard example set $\mathbf{H}$.

3 Experiments

We collected online-class video recordings from a third-party educational platform. Similar to [23, 11], audio tracks are extracted from video recordings and then cut into utterances by a self-trained VAD model [20]. After that, utterances are transcribed into text using a self-trained ASR model [26] with a character error rate (CER) of 11.36% in classroom scenarios. The training set contains 16174 instances and 4088 instances in the validation set. Performance on each category (except *others*) is separately evaluated on a binary test set containing 2000 positive instances that belong to this category, and 2000 negative ones from the other categories (other seven categories of instructions, or *others*).

We select a series of widely-used baselines, including BiLSTM [8], TextRCNN [13], and pre-trained language models: BERT [5], ELECTRA [4], NEZHA [22], RoBERTa [15], and XLNet [24]. Moreover, we compare different strategies of negative example selection in our multi-task framework: (1) random selection from all the instances of other categories, i.e., *M-RoBERTa-All*; and (2) hard example mining, i.e., *M-RoBERTa-Hard*.

3.1 Results Discussion

From Table 1, we can find that pre-trained language models such as ELECTRA, NEZHA, and RoBERTa achieve higher performance than classic approaches,

Table 1. Performance of different pre-trained language models.

Instruction	Model	Accuracy	F1	Instruction	Model	Accuracy	F1
	BiLSTM	0.781	0.783		BiLSTM	0.781	0.791
	TextRCNN	0.785	0.788		TextRCNN	0.785	0.789
	BERT	0.781	0.787		BERT	0.781	0.778
macro-average	ELECTRA	0.791	0.790	micro-average	ELECTRA	0.791	0.794
	NEZHA	0.797	0.803		NEZHA	0.797	**0.797**
	XLNet	0.770	0.775		XLNet	0.770	0.764
	RoBERTa	**0.799**	**0.812**		RoBERTa	**0.799**	0.795

Table 2. Performance of the proposed method and its variants.

Instruction	Model	Accuracy	F1	Instruction	Model	Accuracy	F1
	RoBERTa	0.828	0.831		RoBERTa	0.809	0.829
commending	M-RoBERTa-All	0.831	0.844	guidance	M-RoBERTa-All	0.847	0.850
	M-RoBERTa-Hard	**0.842**	**0.855**		M-RoBERTa-Hard	**0.868**	**0.872**
	RoBERTa	0.803	0.829		RoBERTa	0.788	0.803
summarization	M-RoBERTa-All	0.862	0.875	greeting	M-RoBERTa-All	0.791	0.810
	M-RoBERTa-Hard	**0.876**	**0.886**		M-RoBERTa-Hard	**0.802**	**0.830**
	RoBERTa	0.814	0.830		RoBERTa	0.690	0.725
note-taking	M-RoBERTa-All	0.735	0.771	repeating	M-RoBERTa-All	0.749	0.774
	M-RoBERTa-Hard	**0.886**	**0.889**		M-RoBERTa-Hard	**0.750**	**0.776**
	RoBERTa	0.796	0.787		RoBERTa	0.868	0.859
reviewing	M-RoBERTa-All	**0.824**	**0.811**	example-giving	M-RoBERTa-All	0.861	0.854
	M-RoBERTa-Hard	0.822	**0.811**		M-RoBERTa-Hard	**0.929**	**0.893**
	RoBERTa	0.799	0.812		RoBERTa	0.799	0.795
macro-average	M-RoBERTa-All	0.812	0.824	micro-average	M-RoBERTa-All	0.812	0.804
	M-RoBERTa-Hard	**0.847**	**0.852**		M-RoBERTa-Hard	**0.847**	**0.823**

i.e., BiLSTM and TextRCNN, which indicates their stronger capacity to model dialogic instructions by utilizing contextual information. ELECTRA, RoBERTa, and NEZHA have a higher overall performance than BERT, which is reasonable since they are pre-trained with improved training objectives and larger corpus.

We demonstrate the effectiveness of our multi-task framework by comparing with RoBERTa model trained with a single task i.e., instruction classification. Table 2 shows that: (1) by adding a contrastive loss to enlarge the margin between different categories, *M-RoBERTa-All* outperforms the original RoBERTa model in 6 out of 8 types of dialogic instructions and the overall performance; and (2) by fully utilizing instances misclassified by the model, *M-RoBERTa-Hard* outperforms *M-RoBERTa-All* and achieves the best prediction performance compared with other methods in terms of accuracy, macro- and micro-F1 scores.

4 Conclusion

In this work, we present a multi-task dialogic instruction detection framework using pre-trained language models. Furthermore, we design a strategy to select hard instances and exploit them when training. Experiments conducted on a real-world data set show that our framework outperforms both classic methods and pre-trained language models fine-tuned solely with the classification objective.

Acknowledgment

This work was supported in part by National Key R&D Program of China, under Grant No. 2020AAA0104500 and in part by Beijing Nova Program (Z201100006820068) from Beijing Municipal Science & Technology Commission.

References

1. AN, S.: Capturing the chinese way of teaching: The learning-questioning and learning-reviewing instructional model. In: How Chinese learn mathematics: Perspectives from insiders, pp. 462–482. World Scientific (2004)
2. Anthony, G., Hunter, J., Hunter, R.: Supporting prospective teachers to notice students' mathematical thinking through rehearsal activities. Mathematics Teacher Education and Development **17**(2), 7–24 (2015)
3. Blanchard, N., Brady, M., Olney, A.M., Glaus, M., Sun, X., Nystrand, M., Samei, B., Kelly, S., D'Mello, S.: A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In: International conference on artificial intelligence in education. pp. 23–33. Springer (2015)
4. Clark, K., Luong, M., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generators. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=r1xMH1BtvB
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
6. Dweck, C.S.: Boosting achievement with messages that motivate. Education Canada **47**(2), 6–10 (2007)
7. Goodenow, C.: The psychological sense of school membership among adolescents: Scale development and educational correlates. Psychology in the Schools **30**(1), 79–90 (1993)
8. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. Ieee (2013)
9. HAGHVERDİ, H., Biria, R., Karimi, L.: Note-taking strategies and academic achievement. Dil ve Dilbilimi Çalışmaları Dergisi **6**(1) (2010)
10. Henderlong, J., Lepper, M.R.: The effects of praise on children's intrinsic motivation: A review and synthesis. Psychological Bulletin **128**(5), 774 (2002)
11. Huang, G.Y., Chen, J., Liu, H., Fu, W., Ding, W., Tang, J., Yang, S., Li, G., Liu, Z.: Neural multi-task learning for teacher question detection in online classrooms. In: International Conference on Artificial Intelligence in Education. pp. 269–281. Springer (2020)
12. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop. vol. 2. Lille (2015)
13. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Bonet, B., Koenig, S. (eds.) Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. pp. 2267–2273. AAAI Press (2015), http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9745
14. Lee, P.L., Lan, W., Hamman, D., Hendricks, B.: The effects of teaching notetaking strategies on elementary students' science learning. Instructional Science **36**(3), 191–201 (2008)

15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
16. Osterman, K.F.: Teacher practice and students' sense of belonging. In: International research handbook on values education and student wellbeing, pp. 239–260. Springer (2010)
17. Rinehart, S.D., Stahl, S.A., Erickson, L.G.: Some effects of summarization training on reading and studying. Reading Research Quarterly pp. 422–438 (1986)
18. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 815–823. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7298682, https://doi.org/10.1109/CVPR.2015.7298682
19. Shafto, P., Goodman, N.D., Griffiths, T.L.: A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cognitive Psychology **71**, 55–89 (2014)
20. Tashev, I., Mirsamadi, S.: Dnn-based causal voice activity detector. In: Information Theory and Applications Workshop (2016)
21. Wang, W., Xu, G., Ding, W., Huang, Y., Li, G., Tang, J., Liu, Z.: Representation learning from limited educational data with crowdsourced labels. TKDE (2020)
22. Wei, J., Ren, X., Li, X., Huang, W., Liao, Y., Wang, Y., Lin, J., Jiang, X., Chen, X., Liu, Q.: Nezha: Neural contextualized representation for chinese language understanding. arXiv preprint arXiv:1909.00204 (2019)
23. Xu, S., Ding, W., Liu, Z.: Automatic dialogic instruction detection for k-12 online one-on-one classes. In: International Conference on Artificial Intelligence in Education. pp. 340–345. Springer (2020)
24. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 5754–5764 (2019), https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html
25. Yelland, N., Masters, J.: Rethinking scaffolding in the information age. Computers & Education **48**(3), 362–382 (2007)
26. 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, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018. pp. 5869–5873. IEEE (2018). https://doi.org/10.1109/ICASSP.2018.8461404, https://doi.org/10.1109/ICASSP.2018.8461404