**ABSTRACT**

Online dialogic instructions are a set of pedagogical instructions used in real-world online educational contexts to motivate students, help understand learning materials, and build effective study habits. In spite of the popularity and advantages of online learning, the education technology and educational data mining communities still suffer from the lack of large-scale, high-quality, and well-annotated teaching instruction datasets to study computational approaches to automatically detect online dialogic instructions and further improve the online teaching effectiveness. Therefore, in this paper, we present a dataset of online dialogic instruction detection, **DialogID**, which contains 30,431 effective dialogic instructions. These teaching instructions are well annotated into 8 categories. Furthermore, we utilize the prevalent pre-trained language models (PLMs) and propose a simple yet effective adversarial training learning paradigm to improve the quality and generalization of dialogic instruction detection. Extensive experiments demonstrate that our approach outperforms a wide range of baseline methods. The data and our code are available for research purposes from: https://github.com/ai4ed/DialogID.

**CCS CONCEPTS**

- Applied computing → Computer-managed instruction; E-learning; Interactive learning environments; Computer-assisted instruction.

**KEYWORDS**

dialogic instruction; teaching effectiveness; instruction detection

---

1 INTRODUCTION

The Covid-19 pandemic has brought tremendous changes to educational institutions around the world. With the recent development of technology such as digital video processing and live streaming, various forms of online learning tools emerge and a large number of offline institutions switch to the online mode [6, 16, 18]. In spite of the advantages of online classes and a variety of support from online teaching software, teaching online classes still remains a very challenging task for the well-trained offline classroom instructors. When sitting in front of a camera or a laptop, traditional classroom instructors lack effective pedagogical instructions to ensure the overall quality of their online classes.

Dialogic instructions for online classes promote interactions between teachers and students instead of teacher-presentation only. Besides, they improve the learning interest and the confidence of the student, constructing effective learning habits. Hence, a computational approach to automatically detect the dialogic instructions during the online class seems to provide real-time feedback to teachers and improve their online teaching skills.

However, there are some challenges to build an automatic detection approach of dialogic instruction. Online teaching is not a standardized procedure. Even for the same learning content, the instructors teach varies according to their own pedagogical styles. Furthermore, the different teaching experiences of the instructors also lead to the quality of dialogic instructions. An illustrative example of note-taking\(^1\) instructions is as follows:

- **S1**: Make sure you write down this key point.
- **S2**: Make sure you remember this key point.

\(^1\)Note-taking instructions ask students to take notes of key points.
S1 instruction gives students concrete action, i.e., taking notes, which is helpful for students to build their learning habits. S2 is an ineffective and confusing instruction and doesn’t meet the quality standard of note-taking instruction. An intelligent dialogic instruction detection model is hoped to effectively distinguish these subtle differences and provide instant feedback to online instructors.

Existing educational research has revealed the significance of dialogic instructions on students’ social emotional well-being [29], motivation [13], and academic achievements [8, 20]. Class observation frameworks have been established such as CLASS [22] and COPUS [27]. However, these methods heavily rely on human efforts like manual video coding [23, 25], and hence fail to provide automatic in-time feedback to instructors. Machine learning models are able to learn from human-coded data then make predictions automatically. For example, Donnelly et al. utilized Naive Bayes models able to learn from human-coded data then make predictions automatically. For example, Donnelly et al. utilized Naive Bayes models able to learn from human-coded data then make predictions automatically.

However, even though the aforementioned research focuses on detecting and studying teachers’ dialogic instructions, none of them open source their research datasets. Furthermore, the majority of research works are undertaken in the traditional offline classrooms and their methodologies and paradigms are not applicable to the online learning environments. Therefore, in this work, to help and promote research and development of tasks for online dialogic instruction detection, we present DIALOGID, a high-quality dialogic instruction dataset for improving online teaching effectiveness. DIALOGID contains 30,431 effective dialogic instructions extracted from real-world K-12 online classes. To the best of our knowledge, DIALOGID is one of the first publicly available dialogic instruction datasets collected from online classrooms. Furthermore, we propose a simple yet effective adversarial training (AT) paradigm with pre-trained language models (PLMs) learned from DIALOGID to solve the dialogic instructions detection problem automatically. Experimental results demonstrate the usage and effectiveness of the DIALOGID dataset and the proposed instructions detection approach.

2 DATASET

2.1 Dialogic Instructions

In this work, following many existing pedagogical studies [1, 2, 8, 9, 12, 13, 15, 21, 24, 26, 33], we focus on online dialogic instructions with the following aspects: (1) motivate students and make them feel easy about the class: greeting [9, 21] and commending [8, 13]; (2) help students understand learning materials and retain them: guidance [33], example-giving [26], repeating [2], and reviewing [1]; and (3) build effective learning habits: note-taking [12, 15] and summarization [24].

Therefore we aim to capture these 8 kinds of effective instructions. The definitions and examples of instructions are shown in Table 1. Please note that the scope of dialogic instructions in our work is a superset of the previous study [31].

2.2 Data Annotation

To ensure the annotation quality and the trained AI driven detection models are able to be deployed into the real production systems without any human intervention, we design a 3-step online dialogic instruction annotation process that aims to automatically identify teaching instructions from the entire online classroom recordings. The 3-step process is described as follows.

Step 1: Extract Teacher Utterances. Similar to [14, 31], we extract teacher utterances from the online classroom video recordings and filter out background noises and silence fragments via an in-house voice activity detection (VAD) model. Similar to [28], the in-house VAD model is a four-layer deep neural network trained on online classroom audio data. Please note that there are no voice overlaps as audio recordings of each teacher and student are recorded separately.

Step 2: Generate Dialogic Instruction Candidates. Dialogic instructions only constitute a small portion of teacher utterances within an online course. To make the annotation efficiently and economically, we identify utterance candidates that may contain dialogic instructions. Specifically, we first transcribe each teacher utterance (obtained from Step 1) via a self-trained automatic speech recognition (ASR) model, which is a deep feed-forward sequential memory network to transfer the voice utterances into text information [34]. The ASR model is trained on classroom specific datasets and has a character error rate of 11.36% in the classroom scenarios. Then for each type of dialogic instruction listed in Table 1, we pre-define a list of keywords and use the keyword matching method to find candidate utterances of dialogic instructions. Only the utterance whose transcription is matched with at least one of the keywords will be kept. The pre-defined keywords are constructed by analyzing thousands of online class videos and surveying hundreds of instructors, students, parents and educators. For example, words or phrases like “Hello/Good Morning/Goodbye” and “as seen in/as shown in” are keywords for the greeting and summarization dialogic instructions respectively.

Step 3: Extract Segment-Level Audios for Utterance-Level Annotation. Individual utterance candidates from Step 2 may simply contain one or two sentences that are difficult to annotate due to the lack of classroom contexts. Therefore, to make sure our crowdsourced labels are reliable, we assemble the target utterance candidate, its preceding n utterances and its n following utterances into an audio segment. The crowd workers assign labels after listening to each audio segment. A teacher’s utterance will be labeled as “others” if it doesn’t belong to the 8 categories in Table 1.

2.3 Data Analysis

Dialogic instructions in DIALOGID are collected and constructed from the K-12 online classes at TAL Education Group[^2]. Through the 3-step annotation procedure, we end up with 51,908 annotated samples and 30,431 of them are effective online dialogic instructions. The detailed per type instruction distribution and the corresponding sizes of training, validation, test sets are shown in Table 2. Furthermore, the length distribution (in words) of each type of dialogic instruction is depicted in Figure 1. As we can see, the amounts of different types of dialogic instructions are relatively balanced in DIALOGID and most of the dialogic instructions are short sentences with less than 20 words.

[^2]: TAL Education Group (NYSE:TAL) is an educational technology company dedicated to supporting public and private education across the world.
Table 1: Definitions and examples of dialogic instructions. others contains instructions that are either ineffective or irrelevant.

| Instruction     | Definition                                                                 | Example(s)                                   |
|-----------------|-----------------------------------------------------------------------------|----------------------------------------------|
| commending      | Commending instructions that praise and encourage students.                 | Good job!                                    |
| guidance        | Guiding students to solve a problem step by step.                          | What would happen then?                      |
| summarization    | Wrapping up the lesson or summarizing the content just learned.            | Let’s conclude what we have learned today.   |
| greeting        | Greetings at the beginning of a class; Instructions that help manage the   | How is it going?                             |
|                 | teaching procedures.                                                        | Can you see the slides?                      |
| note-taking      | Instructions that ask students to take notes of key points.                | Make sure you write down this key point.    |
| repeating       | Requiring students to rehearse the content.                                | Could you repeat it?                         |
| reviewing       | Reminding the students what they learned in a previous class.              | Could you remember the words you learned last week? |
| example-giving  | Demonstrating the content by concrete facts.                               | Here is an example.                          |
| others          | Ineffective instructions, or instructions unrelated to the class.          | It’s good weather today.                    |

Table 2: Data statistics of the DIALOGID dataset.

| Instruction     | Train | Validation | Test | Total | AVG Len | STD Len |
|-----------------|-------|------------|------|-------|---------|---------|
| commending      | 2,437 | 320        | 692  | 3,449 | 17.2    | 16.3    |
| guidance        | 2,987 | 425        | 881  | 4,293 | 23.4    | 18.4    |
| summarization    | 2,206 | 307        | 588  | 3,101 | 27.1    | 23      |
| greeting        | 1,798 | 243        | 529  | 2,570 | 15.5    | 13.8    |
| note-taking      | 2,667 | 394        | 782  | 3,843 | 19.9    | 15.3    |
| repeating       | 2,488 | 368        | 705  | 3,561 | 19.9    | 14.2    |
| reviewing       | 2,793 | 402        | 786  | 3,881 | 27.3    | 19.6    |
| example-giving  | 3,977 | 550        | 1,106 | 5,633 | 24.8    | 20.5    |
| others          | 14,982| 2,182      | 4,313| 21,477| 21.6    | 17.1    |
| Total           | 36,335| 5,191      | 10,382| 51,908| 22      | 18      |

3 AN ADVERSARIAL TRAINING ENHANCED DETECTION FRAMEWORK

In this section, we describe our dialogic instruction detection framework, which has two key components: (1) a FLMM, which serves as the base model in the classification task; and (2) an adversarial training learning module, which improves the model generalization from the very limited and noisy teacher instruction transcriptions.

3.1 Pre-trained Language Models

Traditional machine learning models use word vectors as inputs, which are static and not able to extract contextual information. By contrast, more recent PLMs learn contextual embeddings with their Transformer-based architectures. Therefore in this study we utilize PLMs as our base model in our detection framework.

To perform the instruction detection task on a sentence \( x = (x_1, \cdots, x_n) \) which contains \( n \) tokens, similar to [5, 17], we first add a special token \([CLS]\) in front of the sentence. After that, embeddings of each token \((E_{[CLS]}, E_1, \cdots, E_n)\) are fed into multiple Transformer encoders sequentially, where each token gradually captures contextual information of the sentence. Finally at the last layer of Transformer encoders, hidden states of each token are extracted, and the hidden state of the special token \([CLS]\) is treated as the representation of the sentence.

In our study, we utilize RoBERTa [17] pre-trained model, which is a Transformer-based model sharing the same structure with BERT [5], while several improvements are made at the pre-training stage, including removing the next-sentence prediction objective of BERT, and using dynamic masking, etc. We also experiment with other recently proposed PLMs and details are discussed in Section 4.

3.2 Adversarial Training Module

Adversarial training is an efficient regularization technique that provides an efficient way to not only improve the robustness of the DNNs against perturbations but also enhance its generalization over original inputs by training DNNs to correctly classify both of the original inputs and adversarial examples (AEs) [10, 11, 19]. Similar to the pioneer work of Miyato et al. who extended AT to text classification [19], we create AEs by adding adversarial perturbations on intermediate representations in the embedding layers and use AEs to optimize the model parameters for better generalization. Specifically, the adversarial perturbation \( \epsilon \) is computed by an efficient and fast gradient approximation method developed by Goodfellow et al. [10] as follows:

\[
x' = x + \epsilon; \quad \epsilon = \frac{g}{\|g\|_2}; \quad g = \nabla_x L(x, \theta)
\]

where \( x' \) and \( x \) denote the perturbed and original representations in neural networks’ embedding layers respectively. \( \epsilon \) is a hyperparameter to control the norm of perturbations and \( \theta \) is the model parameters.

4 EXPERIMENTS

To comprehensively assess DIALOGID and the proposed method, except for RoBERTa, we additionally select a series of widely-used text classification models, including BERT [5], ELECTRA [3], MacBERT [4], and XLNet [32]. Moreover, we conduct an ablation study to demonstrate the performance improvement enhanced by the adversarial training module. The proposed AT enhanced approach is denoted as “RoBERTa+AT” in the following section. Please note that the AT module can be incorporated with any PLM. PLMs for our experiments can be found in this repository 3. For each model, we set max_len to 128 and learning rate to 1e-5. The number of epochs is set to 100, and we early stop in the training phase if the model doesn’t get improvement in validation datasets for 5 epochs.

3https://github.com/ymcui/Chinese-BERT-wwm
4.1 Results

Prediction with PLMs. We compare the performance of different PLMs in terms of precision, recall and F1 scores. Results are shown in Table 3 (per type) and Table 4 (overall). When comparing RoBERTa to other PLMs such as BERT, ELECTRA, MacBERT, and XLNet, we can find that RoBERTa achieves the highest prediction performance than other approaches in terms of F1 score, which indicates their stronger capacity to model dialogic instructions. When looking into each category, interestingly we can observe that RoBERTa does not always achieve the top performance. In detail, they show inferior performance compared with BERT and MacBERT in summarization, greeting and repeating, which have the least samples compared to instructions in other categories.

Table 3: Prediction performance per instruction type of all different baselines in terms of precision, recall and F1 score.

| Dialogic Instruction | Model  | Precision  | Recall  | F1    |
|----------------------|--------|------------|---------|-------|
| commending           | BERT   | 0.8274     | 0.8801  | 0.8529|
|                      | ELECTRA| 0.8093     | 0.8829  | 0.8445|
|                      | MacBERT| 0.8219     | 0.8801  | 0.8500|
|                      | XLNet  | 0.8343     | 0.8223  | 0.8282|
|                      | RoBERTa| 0.8013     | 0.9263  | 0.8592|
|                      | RoBERTa+AT| 0.8083   | 0.9263  | 0.8633|
| summarization        | BERT   | 0.9039     | 0.8963  | 0.9001|
|                      | ELECTRA| 0.8542     | 0.9167  | 0.8843|
|                      | MacBERT| 0.8882     | 0.9184  | 0.9030|
|                      | XLNet  | 0.8938     | 0.8878  | 0.8908|
|                      | RoBERTa| 0.8834     | 0.9150  | 0.8989|
|                      | RoBERTa+AT| 0.8834  | 0.9150  | 0.8989|
| note-taking          | BERT   | 0.8100     | 0.9488  | 0.8740|
|                      | ELECTRA| 0.8082     | 0.9373  | 0.8680|
|                      | MacBERT| 0.7940     | 0.9514  | 0.8656|
|                      | XLNet  | 0.8491     | 0.8632  | 0.8561|
|                      | RoBERTa| 0.8201     | 0.9501  | 0.8803|
|                      | RoBERTa+AT| 0.8493  | 0.8939  | 0.8710|
| reviewing            | BERT   | 0.8162     | 0.9720  | 0.8873|
|                      | ELECTRA| 0.8264     | 0.9644  | 0.8912|
|                      | MacBERT| 0.8284     | 0.9644  | 0.8912|
|                      | XLNet  | 0.8315     | 0.9542  | 0.8886|
|                      | RoBERTa| 0.8346     | 0.9631  | 0.8943|
|                      | RoBERTa+AT| 0.8412  | 0.9567  | 0.8952|

4.2 Qualitative Analysis. We further demonstrate the effectiveness of RoBERTa+AT by visualizing the learned representations, shown in Figure 2. Instances of nine categories in the testing set are fed into the trained models, and their representations i.e., the tensor values of the special token [CLS] are collected. Dimension reduction method t-SNE [30] is performed so that the representations can be visualized in the 2-dimensional space. From the figure, we can see that representations of instances in different categories are well separated by the proposed RoBERTa+AT with significant margins between different categories.

Table 4: Overall prediction performance of different models.

| Model          | Precision  | Recall  | F1    |
|----------------|------------|---------|-------|
| BERT           | 0.8534     | 0.9092  | 0.8795|
| ELECTRA        | 0.8338     | 0.9145  | 0.8720|
| MacBERT        | 0.8534     | 0.9064  | 0.8779|
| XLNet          | 0.8505     | 0.8920  | 0.8701|
| RoBERTa        | 0.8483     | 0.9167  | 0.8602|
| RoBERTa+AT     | 0.8545     | 0.9183  | 0.8649|

5 CONCLUSION

In this work, we introduce DialogID, a dialogic instruction dataset that contains 8 categories of the online class instructions collected from real-world K-12 online classrooms. Experiments conducted on DialogID show the effectiveness and superiority of our proposed approach against a wide range of baselines.

ACKNOWLEDGMENTS

This work was supported in part by National Key R&D Program of China, under Grant No. 2020AAA0104500; in part by Beijing Nova Program (Z20110006820068) from Beijing Municipal Science & Technology Commission and in part by NFSC under Grant No. 61877029.
REFERENCES

[1] Shuha AN. 2004. Capturing the Chinese way of teaching: The learning-questioning and learning-reviewing instructional model. In How Chinese learn mathematics: Perspectives from insiders. World Scientific, 462–482.

[2] Glenda Anthony, Jodie Hunter, and Roberta Hunter. 2015. Supporting Prospective Teachers to Notice Students' Mathematical Thinking through Rehearsal Activities. Mathematics Teacher Education and Development 17, 2 (2015), 7–24.

[3] Kevin Clark. Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In International Conference on Learning Representations.

[4] Yiming Cui, Wanshang Che, Ting Liu, Bing Qin, Shujin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922 (2020).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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). 4171–4186.

[6] Shivangi Dhawan. 2020. Online learning: A panacea in the time of COVID-19 crisis. Journal of Educational Technology Systems 49, 1 (2020), 5–22.

[7] Patrick J Donnelly, Nathan Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystran, and Sidney K D’Mello. 2016. Automatic teacher modeling from live classroom audio. In Proceedings of the 2016 Conference on User Modeling, Adaptation and Personalization. 45–53.

[8] Carol S Dweck. 2007. Boosting achievement with messages that motivate. Education Canada 47, 2 (2007), 6–10.

[9] Carol Goodenow. 1993. The psychological sense of school membership among adolescents: Scale development and educational correlates. Psychology in the Schools 30, 1 (1993), 79–90.

[10] Ian J Goodfellow, Jonathan Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).

[11] Xiaopeng Guo, Zhiwei Huang, Jie Gao, Mingyu Shang, Maojing Shu, and Jun Sun. 2021. Enhancing Knowledge Tracing via Adversarial Training. In Proceedings of the 29th ACM International Conference on Multimedia. 367–375.

[12] Hamid Haghverdi, Reza Biria, and Lotfollah Karimi. 2010. Note-taking strategies and academic achievement. Dil ve Dilbilimi Çalışmaları Dergisi 6, 1 (2010).

[13] Jennifer Henderlong and Mark R Lepper. 2002. The effects of praise on children’s intrinsic motivation: A review and synthesis. Psychological Bulletin 128, 5 (2002), 774.

[14] Gao Yan Huang, Jiahao Chen, Haochen Liu, Weiping Fu, Wenbiao Ding, Jiliang Tan, Songfanyang Yang, Gusliang Li, and Zitao Liu. 2020. Neural multi-task learning for teacher question detection in online classrooms. In International Conference on Artificial Intelligence in Education. Springer, 269–281.

[15] Pai-Lin Lee, Wållén Lång, Douglas Hamman, and Bret Hendrickson. 2008. The effects of teaching notetaking strategies on elementary students' science learning. Instructional Science 36, 3 (2008), 191–201.

[16] Hang Li, Yu Kang, Wenbiao Ding, Song Yang,Songfanyang Yang, Gao Yan Huang, and Zitao Liu. 2020. Multimodal learning for classroom activity detection. In IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 9234–9238.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019).

[18] Zitao Liu, Guowei Xu, Tianqiao Liu, Weiping Fu, Yabu Qi, Wenbiao Ding, Yujia Song, Chaoyou Guo, Cong Kong, Songfan Yang, et al. 2020. Dolphm: a spoken language proficiency assessment system for elementary education. In Proceedings of The Web Conference 2020. 2641–2647.

[19] Takara Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In International Conference on Learning Representations.

[20] Barbara E Moely, Silvia S Hart, Linda Leal, Kevin A Santalli, Nirmala Rao, Terry Johnson, and Libby Burney Hamilton. 1992. The teacher’s role in facilitating memory and study strategy development in the elementary school classroom. Child Development 63, 3 (1992), 653–672.

[21] Karen F Osterman. 2010. Teacher Practice and Students’ Sense of Belonging. International Research Handbook on Values Education and Student Wellbeing (2010), 239.

[22] Robert C Pianta, Karen M La Paro, and Bridget K Hamre. 2008. Classroom Assessment Scoring System. Manual K-3. Paul H Brookes Publishing.

[23] Anna-Katharina Prasstorius and Charalampos Y Charalambous. 2018. Classroom Observation Frameworks for Studying Instructional Quality: Looking Back and Looking Forward. ZDM: The International Journal on Mathematics Education 50, 3 (2018), 535–553.

[24] Steven D Rinehart, Steven A Stahl, and Lawrence G Erickson. 1986. Some effects of summarization training on reading and studying. Reading Research Quarterly (1986), 422–438.

[25] Barak Rosenshine. 2012. Principles of instruction: Research-based strategies that all teachers should know. American Educator 36, 1 (2012), 12.

[26] Patrick Shafio, Noah D Goodman, and Thomas L Griffiths. 2014. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cognitive Psychology 71 (2014), 55–89.

[27] Michelle K Smith, Francis HM Jones, Sarah L Gilbert, and Carl E Wieman. 2013. The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. CBE—Life Sciences Education 12, 4 (2013), 618–627.

[28] Ivan Tashev and Seyeddinasab Mirmasoud. 2016. DNN-based causal voice activity detector. In Information Theory and Applications Workshop.

[29] Jaclyn E Tennant, Michelle K Demaray, Christine K Malecki, Melissa N Terry, Michael Clary, and Nathan Elzinga. 2015. Students’ ratings of teacher support and academic and social–emotional well-being. School Psychology Quarterly 30, 4 (2015), 494.

[30] L Van, der Maaten and Ge Hinton. 2008. (Visualizing High-Dimensional Data Using t-SNE). Journal of Machine Learning Research 9, 2 (2008), 2579–2605.

[31] Shiting Xu, Wenbiao Ding, and Zitao Liu. 2020. Automatic dialogic instruction instruction age. In International Conference on Artificial Intelligence in Education. Springer, 494–505.

[32] Zhihui Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems 32 (2019).

[33] Neola Ielly and Jennifer Masters. 2007. Rethinking scaffolding in the information age. Computers & Education 48, 3 (2007), 362–382.

[34] Shiliang Zhang, Ming Lei, Zhiye Yan, and Lirong Dai. 2018. Deep-FSMN for large vocabulary continuous speech recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 5869–5873.