Improving Contextual Language Models for Response Retrieval in Multi-Turn Conversation

Junyu Lu [1*] Xiancong Ren [1*] Yazhou Ren [1] Ao Liu [1] Zenglin Xu [2,3,1⋆]
[1]SMILE Lab, Sch. Computer Science and Engineering, University of Electronic Science and Technology of China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[3]Artificial Intelligence Center, Peng Cheng Lab, Shenzhen, China
{cs.junyu,renxiancong}@gmail.com,yazhou.ren@uestc.edu.cn,{zeitmond,zenglin}@gmail.com

ABSTRACT

As an important branch of current dialogue systems, retrieval-based chatbots leverage information retrieval to select proper pre-defined responses. Various promising architectures have been designed for boosting response retrieval, however, few researches exploit the effectiveness of the pre-trained contextual language models. In this paper, we propose two approaches to adapt contextual language models in dialogue response selection task. In detail, the *Speaker Segmentation* approach is designed to discriminate different speakers to fully utilize speaker characteristics. Besides, we propose the *Dialogue Augmentation* approach, *i.e.*, cutting off real conversations at different time points, to enlarge the training corpora. Compared with previous works which use utterance-level representations, our augmented contextual language models are able to obtain top-hole contextual dialogue representations for deeper semantic understanding. Evaluation on three large-scale datasets has demonstrated that our proposed approaches yield better performance than existing models.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**.

KEYWORDS

Response Retrieval, Pre-trained Language Model, Augmentation

ACM Reference Format:
Junyu Lu, Xiancong Ren, Yazhou Ren, Ao Liu, Zenglin Xu. 2020. Improving Contextual Language Models for Response Retrieval in Multi-Turn Conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401255

1 INTRODUCTION

Dialogue systems have been widely used in a variety of applications, spanning from entertainment and personal assistance, to customer services. The key problem for building such a dialogue system is how to reply to a message with a proper (i.e., human-like and natural) response. Roughly speaking, existing approaches are either retrieval-based or generation-based. Unlike generation-based approaches which generate the entire responses, retrieval-based approaches leverage information retrieval (IR) techniques to rank a pool of pre-built candidates and select a proper response from the top ranked ones. Obviously, retrieval-based chatbots have the advantages of providing fluent and informative responses, and thus have attracted significant attention in the information retrieval community [5, 13, 14].

Early works narrowly consider hand-crafted rules in single-turn conversation. Despite their significance, these systems neglect the contexts of the historical conversation session which play a pivotal role in the following chat. Along with the prosperity of deep learning, multi-turn neural conversation systems have been devised, where textual information is represented as a dense and continuous vector. These models could be classified into two categories: (1) representation-focused models [7, 13, 17], which firstly learn the representations of query and document separately, then measure the matching degree between them. (2) interaction-focused models [8, 10–12, 15, 16, 18], which build a matching module to compute semantic interaction matrices between query and document, then the interaction matrices will be fed into CNNs to obtain the similarity score. Nevertheless, they still lack enough semantic comprehension of human conversations. More recently, high-capacity contextual language models, like BERT [2], lead to significant improvement on downstream natural language understanding tasks, such as Question Answering (QA) and Natural Language Inference (NLI) [2]. Yet it remains a worthy question how to incorporate pre-trained contextual language models into multi-turn retrieval-based chatbots.

In this paper, we study the effectiveness of three pre-trained language models, BERT, BERT_{WWM} [1], and RoBERTa [6] for retrieval-based chatbots and propose two adaptation methods to enhance their performance. BERT_{WWM} is an improved variant of BERT, which leverages Whole Word Masking technique to mask a complete word in pre-training rather than simply mask WordPiece tokens. RoBERTa further improves BERT by removing the next sentence prediction objective and dynamically changing the masking pattern during pre-training. To fit these BERT-like models with multi-turn dialogues, it is natural to follow the sequence pair setting to fine-tune BERT which is commonly adopted in QA tasks [2]. However, such sequence pair inputs may mislead BERT in the ongoing conversation since BERT can not understand which speaker

*Both authors contributed equally to this research. ⋆Corresponding author.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401255

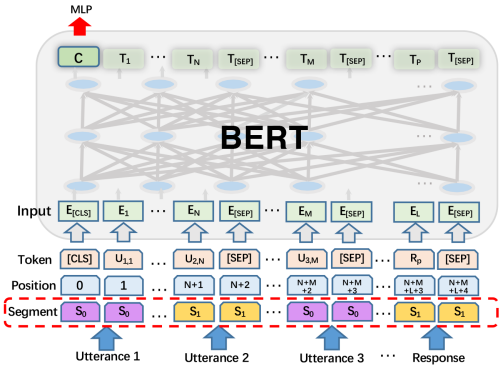

Figure 1: BERT with Speaker Segmentation. The input embeddings are the sum of the corresponding token, position, and segment embeddings. The segment embeddings, S_0 and S_1, represent different speakers respectively. "[CLS]" is a classification token that used to capture an aggregate hidden representation (C) of the entire sequence.

should be replied to. Motivated by this, we propose *Speaker Segmentation* to discriminate different speakers. By using this scheme, BERT-like matching models can understand who the next speaker is and utilize this prior information to select speaker-related response.

Moreover, most of the current available corpora in response retrieval [7] just adopt a simple heuristic to automatically construct a training set where negative candidates are randomly sampled. Last years, Feng et al. [3] and Li et al. [4] proved that random negative responses bring strong noise corruption in response retrieval datasets. In fact, multi-turn dialogues have some important characteristics in terms of consistency and logicality. To fully exploit such characteristics, we propose a *Dialogue Augmentation* strategy to obtain sufficient positive and negative responses for training which are directly sampled from parts of the original dialogues.

To the best of our knowledge, this paper is the first to utilize diverse pre-trained language models in retrieval-based chatbots. Besides, we experiment three different pre-trained language models and we are the first to improve them via *Speaker Segmentation* and *Dialogue Augmentation* strategies. Furthermore, intensive experiments show that our methods can achieve state-of-the-art results on large scale datasets. Ablation study also shows the effectiveness of our adaptation methods. The code and reproducible experimental detail are publicly available at https://github.com/CSLujunyu/Improving-Contextual-Language-Modelsfor-Response-Retrieval-in-Multi-Turn-Conversation.

2 METHODOLOGY

Our work mainly follows BERT architecture described by Devlin et al. [2]. We apply two new approaches to BERT for response retrieval task: a **Speaker Segmentation** scheme to match speaker-related responses, and **Dialogue Augmentation** for better model generalization. At the last layer, the "[CLS]" representation is fed into a Multi-Layer Perceptron (MLP) to identify whether the response is proper or not. The overall model architecture is illustrated in Figure 1.

2.1 Speaker Segmentation

For the tasks involving a two-sequence input, such as QA (<Question,Answer>) and IR (<Query,Document>), a common input pattern of BERT is usually formulated as "[CLS] Sequence_A [SEP] Sequence_B [SEP]", to which we refer as sequence pairs (SP). In IR tasks, for example, the SP format is "[CLS] Query [SEP] Document [SEP]". Thus it is straightforward to adopt SP format in dialogue tasks by the following steps: (1)concatenating multiple utterances into a context sequence; (2)packing the context sequence and a response together with a separation token ([SEP]); (3)measuring their matching degree.

Furthermore, we propose *Speaker Segmentation* (SS) to extend BERT into a multi-turn manner. As shown in Figure 1, we denote the end of utterances with "[SEP]" to separate utterances of different speakers. The input representation is constructed by summing the corresponding token, position, and segment embeddings, out of which, the segment embeddings is initialized with respect to the corresponding speaker. In a nutshell, people usually take turns expressing their ideas in two-person conversation so that we can simply calculate the segment embeddings via $S_{k\%2}$, where $\{S_0, S_1\}$ is the segment embeddings in pre-trained BERT and k is the sequential number of current utterance. Additionally, the segment embedding of "[CLS]" is always initialized as S_0 [2] while "[SEP]" depends on the corresponding utterance. By denoting u^T as the T-th utterance, and $\mathcal{R}$ as a response candidate, we illustrate the formats of SP and SS as follows:

$$\begin{aligned} \text{SP:}\quad & \text{[CLS]}\ u^0\ u^1\ \ldots\ u^T\ \text{[SEP]}\ \mathcal{R}\ \text{[SEP]}, \\ \text{SS:}\quad & \text{[CLS]}\ u^0\ \text{[SEP]}\ u^1\ \text{[SEP]}\ \ldots\ u^T\ \text{[SEP]}\ \mathcal{R}\ \text{[SEP]}. \end{aligned}$$

2.2 Dialogue Augmentation

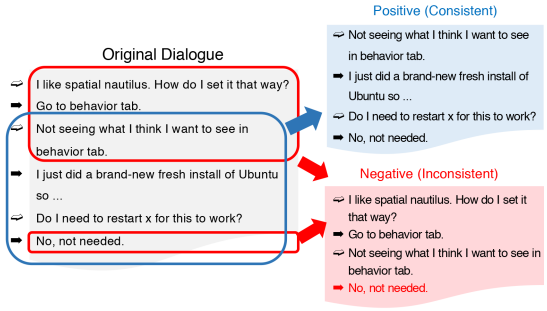

Figure 2: Dialogue Augmentation. The consistent part of dialogue inside the "blue" box can be seen as an extra positive sample, while the inconsistent "red" part is negative.

To improve the quality and quantity of training samples, we propose a *Dialogue Augmentation* (DA) method to provide sufficient conversations for training. Inspired by the construction of Advising Dialogue dataset [1], we use a similar sampling scheme to generate additional data from the original dialogues. However, since BERT-like models focus on learning contextual representations

[1]Dialog System Technology Challenges (DSTC)

Model	Ubuntu				Douban						E-commerce		
	R_2@1	R_{10}@1	R_{10}@2	R_{10}@5	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
TF-IDF [7]	0.659	0.410	0.545	0.708	0.331	0.359	0.180	0.960	0.172	0.405	0.159	0.256	0.477
BM25-PRF [9]	-	0.528	0.655	0.829	-	-	-	-	-	-	-	-	-
DL2R [13]	0.899	0.626	0.783	0.944	0.488	0.527	0.330	0.193	0.342	0.705	0.399	0.571	0.842
Multi-View [17]	0.908	0.662	0.801	0.951	0.505	0.543	0.342	0.202	0.350	0.729	0.421	0.601	0.861
SMN [12]	0.926	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
DUA [16]	-	0.752	0.868	0.962	0.551	0.599	0.421	0.243	0.421	0.780	0.501	0.700	0.921
DAM [18]	0.938	0.767	0.874	0.969	0.550	0.601	0.427	0.254	0.410	0.757	-	-	-
MRFN [10]	0.945	0.786	0.886	0.976	0.571	0.617	0.448	0.276	0.435	0.783	-	-	-
IoI [11]	0.947	0.796	0.894	0.974	0.573	0.621	0.444	0.269	0.451	0.786	0.563	0.768	0.950
MSN [15]	-	0.800	0.899	0.978	0.587	0.632	0.470†	0.295†	0.452	0.788	0.606	0.770	0.937
$BERT_{BASE}$ [2]	0.939	0.783	0.877	0.968	0.563	0.599	0.407	0.253	0.437	**0.796**	0.532	0.726	0.943
$BERT_{BASE}$-SS-DA	**0.953**	**0.813**	**0.901**	0.977	**0.590**	0.627	0.437	0.266	**0.470**	**0.841**	**0.623**	**0.819**	**0.964**
BERT-WWM_{BASE}-SS-DA	-	-	-	-	**0.602**†	**0.643**	0.458	0.280	**0.491**	**0.843**	**0.648**†	**0.843**†	**0.980**†
RoBERTa$_{BASE}$-SS-DA	**0.955**†	**0.826**†	**0.909**†	**0.978**†	**0.602**†	**0.646**†	0.460	0.280	**0.495**†	**0.847**†	**0.627**	**0.835**	**0.980**†

Table 1: Comparison of different models over the Ubuntu, Douban, and E-commerce testsets. Numbers in bold font mean the better result compared with the baselines. † implies the best in each column.

from the context, a coherent context dialogue is more suitable in our case. Therefore, instead of randomly sampling from multiple turns of a dialogue as Advising data, we extract coherent parts of dialogues from different time span. As illustrated in Figure 2, we randomly slice long conversations at different turns to derive partial conversations as positives; while informative negative samples are constructed by removing some intermediate turns of conversation. The last turn of augmented samples can be seen as a positive or negative response. It is also worth noting that this kind of augmentation methods may cause severe overfitting in previous non-BERT models due to their repeatability in chat context, however, pre-trained contextual language models can to some extent avoid such problems. This is because most of existing works independently encode each utterance, while BERT-like models can well capture the contextual information from the entire dialogue to avoid the influence of similar samples.

3 EXPERIMENT

3.1 Datasets and Baselines

We conduct experiments upon three standard response selection datasets: **(1)Ubuntu Dialogue Corpus** includes English multi-turn dialogues extracted from the Ubuntu chat logs about Ubuntu-related technical support. **(2)Douban Conversation Corpus** consists of Chinese human-human conversations crawled from the Douban group [3] on open-domain topics, such as movies, books, and etc. **(3)E-commerce Corpus** contains Chinese real world conversations between customers and customer service staff in Taobao. And we compare our approaches with statistic baselines (**TF-IDF** and **BM25-PRF** [9]) and current neural architectures, including Deep Learning-to-Respond (**DL2R**) [13], **Multi-view** matching model [17], Sequential Matching Network (**SMN**) [12], Deep Utterance Aggregation (**DUA**) [16], Deep Attention Matching network (**DAM**) [18], Multi-Representation Fusion Network (**MRFN**) [10], Interaction-over-Interaction network (**IoI**) [11], and Multi-hop Selector Network (**MSN**) [15]. These statistical and deep neural methods are shown in the first two areas of Table 1 respectively.

[2]We omit the suffix of sequence pair format for brevity.
[3]https://www.douban.com/

3.2 Implementation Details

We denote the experimented models that use *Speaker Segmentation* and *Dialogue Augmentation* with corresponding suffix "-SS" and "-DA". Due to the computation resource limitation, we focus on the setting of fine-tuning the **BASE** pre-trained models (instead of training from scratch). And we do not experiment pre-trained BERT-WWM_{BASE} on the Ubuntu because it is unavailable in English. All of $BERT_{BASE}$ [2], BERT-WWM_{BASE} [1], and RoBERTa$_{BASE}$ [6] follow Huggingface's Pytorch implementations [4]. We use a batch size of 500 and fine-tune for 5 epochs over three datasets. For each dataset, we choose 5e-5 as the initial learning rate and apply Early Stopping to avoid overfitting. We limit the length of the input sequence to 256 words for Ubuntu and Douban (128 words for E-commerce). In addition, Dialogue Augmentation is only applied to conversations longer than 9 turns for the diversity of augmented samples.

3.3 Experimental Results

Table 1 lists the performance of our methods and baselines on standard IR evaluation metrics as the same as prior works [8, 11, 12, 15, 18]. We can see that our proposed methods, denoted in boldface type, outperform most of previous methods by a significant margin. Although P@1 and R_{10}@1 are a little lower in Douban experiment, other higher results such as R_{10}@2 and MAP indicate that our methods can retrieve relevant responses in most cases, better than the state-of-the-art. Moreover, careful hyperparameter tuning may contribute to further improvement.

4 DISCUSSION

4.1 Ablation Study

In this part, we conduct ablation study to investigate the performance of each component of our models. Results are reported in Table 2. We firstly change the input of Sequence Pair with Speaker Segmentation format in $BERT_{BASE}$ architecture. When only employing *Speaker Segmentation*, the R_{10}@1 score is improved by average 4.3%, indicating that fine-tuning BERT with separated speaker-related context is a more suitable way for multi-turn dialogue. Then,

[4]https://github.com/huggingface/transformers

Models	Ubuntu	Douban	E-commerce
$BERT_{BASE}$	0.783	0.253	0.532
+SS	0.811(↑ 3.5%)	0.262(↑ 3.5%)	0.595(↑ 11.8%)
+SS +DA	0.813(↑ 3.8%)	0.266(↑ 5.1%)	0.623(↑ 17.1%)
$RoBERTa_{BASE}$	0.787	0.276	0.585
+SS	0.820(↑ 4.1%)	0.279(↑ 1.0%)	0.599(↑ 2.3%)
+SS +DA	0.826(↑ 4.9%)	0.280(↑ 1.4%)	0.627(↑ 7.1%)

Table 2: Comparison of R_{10}@1 on three testsets. Relative improvements are shown in brackets.

we employ *Speaker Segmentation* and *Dialogue Augmentation* together, which achieves about 6.5% improvement at R_{10}@1. It also shows that current available corpora are the bottleneck of BERT in response retrieval. We conjecture that future superiority may come from more reliable augmentation methods to automatically construct datasets.

4.2 Importance of Pre-training and Fine-tuning

We analyze how much BERT benefits from "Pre-training and Fine-tuning" in Figure 3. In order to investigate the importance of pre-training procedure, we randomly initialize the BERT module instead of using a warm-start one provided by Google [5]. In this setting, BERT without pre-training only achieves 0.234 R_{10}@1 score in the Ubuntu validation, even can not catch up with the baseline. We find it difficult to optimize randomly initialized BERT architecture in downstream response retrieval as the training loss seems to have a convergence problem. Furthermore, we do not fine-tune BERT's parameters and add three extra fully connected layers to classifier. Although it can also boost R_{10}@1 from 0.234 to 0.503, there is still a significant gap to reach the state-of-the-art.

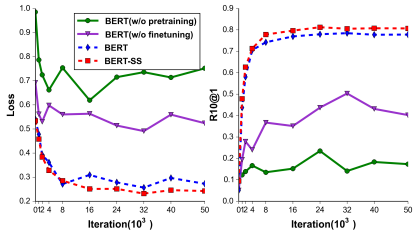

Figure 3: Trends of loss and R_{10}@1 on the Ubuntu dataset.

4.3 Limitation and Future Work

We observe from the bad cases that domain-specific knowledge is essential for future development. For an example in Douban, given a conversation context (A: "Which lyrics of Cai Jianya can touch you?", B: "Dark cloud, go away.", A: "What is the name of this song?"), response ("Don't trouble me.") is logically mismatched because these models do not understand "Don't trouble me" is a Chinese song sung by Cai Jianya. Hence, it would be a promising solution for further improvement by incorporating Knowledge Graph (KG), like music KG, movie KG, etc. It's also worth mentioning that logical information could be essential for the development of future conversation agents since some responses are obscure, subjective, and corrupted.

[5]https://github.com/google-research/bert

5 CONCLUSION

This work studies the pre-trained language models in multi-turn conversation response retrieval. We find that these pre-trained models show great power on matching context and response by providing contextual dialogue representation. Two approaches, *Speaker Segmentation* and *Dialogue Augmentation*, are developed to enhance contextual representation for dialogue modeling. The experiments and further analysis show the superior performance of our approaches over baselines in this task.

ACCKNOWLEDGEMENT

This work was partially supported by the National Key Research and Development Program of China (No. 2018AAA0100204).

REFERENCES

[1] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *CoRR* abs/1906.08101 (2019). arXiv:1906.08101
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805
[3] Jiazhan Feng, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Learning a Matching Model with Co-teaching for Multi-turn Response Selection in Retrieval-based Dialogue Systems. *CoRR* abs/1906.04413 (2019).
[4] Jia Li, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Sampling Matters! An Empirical Study of Negative Sampling Strategies for Learning of Matching Models in Retrieval-based Dialogue Systems. In *EMNLP-IJCNLP'19*. 1291–1296.
[5] Ao Liu, Lizhen Qu, Junyu Lu, Chenbin Zhang, and Zenglin Xu. 2019. Machine Reading Comprehension: Matching and Orders. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. 2057–2060.
[6] 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. *CoRR* abs/1907.11692 (2019).
[7] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL'15*. 285–294.
[8] Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Constructing Interpretive Spatio-Temporal Features for Multi-Turn Responses Selection. In *ACL'19*. 44–50.
[9] S. E. Robertson and S. Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR '94*, Bruce W. Croft and C. J. van Rijsbergen (Eds.). 232–241.
[10] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Multi-Representation Fusion Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *WSDM*. https://doi.org/10.1145/3289600.3290985
[11] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues. In *ACL'19*. 1–11.
[12] Yu Wu, Wei Wu, Ming Zhou, and Zhoujun Li. 2016. Sequential Match Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. *CoRR* abs/1612.01627 (2016). arXiv:1612.01627
[13] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *SIGIR'16*. 55–64. https://doi.org/10.1145/2911451.2911542
[14] Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint Learning of Response Ranking and Next Utterance Suggestion in Human-Computer Conversation System. In *SIGIR '17 (SIGIR '17)*. 685–694. https://doi.org/10.1145/3077136.3080843
[15] Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop Selector Network for Multi-turn Response Selection in Retrieval-based Chatbots. In *EMNLP-IJCNLP'19*. 111–120.
[16] Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. *CoRR* abs/1806.09102 (2018). arXiv:1806.09102
[17] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view Response Selection for Human-Computer Conversation. In *EMNLP'16*. 372–381.
[18] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In *ACL'18*. 1118–1127.