Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots

Jia-Chen Gu¹, Tianda Li², Quan Liu¹³, Zhen-Hua Ling¹, Zhiming Su³, Si Wei³, Xiaodan Zhu²
¹National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China
²ECE & Ingenuity Labs, Queen’s University, Kingston, Canada
³State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Hefei, China

gujc@mail.ustc.edu.cn,tianda.li/xiaodan.zhu@queensu.ca,quanliu/zhling@ustc.edu.cn,zmsu/siwei@iflytek.com

ABSTRACT
In this paper, we study the problem of employing pre-trained language models for multi-turn response selection in retrieval-based chatbots. A new model, named Speaker-Aware BERT (SA-BERT), is proposed in order to make the model aware of the speaker change information, which is an important and intrinsic property of multi-turn dialogues. Furthermore, a speaker-aware disentanglement strategy is proposed to tackle the entangled dialogues. This strategy selects a small number of most important utterances as the filtered context according to the speakers’ information in them. Finally, domain adaptation is performed to incorporate the in-domain knowledge into pre-trained language models. Experiments on five public datasets show that our proposed model outperforms the present models on all metrics by large margins and achieves new state-of-the-art performances for multi-turn response selection.

CCS CONCEPTS
- Information systems → Retrieval models and ranking.

KEYWORDS
Speaker-aware BERT, multi-turn response selection, retrieval-based chatbot

1 INTRODUCTION
Chatbots aim to engage users in open-domain human-computer conversations and are currently receiving increasing attention. The existing work on building chatbots includes generation-based methods and retrieval-based methods. In this paper, we focus on the retrieval-based chats and study the problem of multi-turn response selection. This task aims to select the best-matched response from a set of candidates, given the context of a conversation which is composed of multiple utterances [6, 7, 10].

Previous work has kept utterances separated and performs matching within a representation-interaction-aggregation framework. These methods further extended the matching and attention architectures which improved the performance on this task, such as SMN [10], DAM [13], IMN [3], IoI[9] and MSN [11]. Recently, pre-trained language models have shown to achieve state-of-the-art performance on a wide range of NLP tasks [1]. [4] made the first attempt to employ pre-trained models for response selection. It adopted a simple strategy by concatenating the context utterances and the response literally, and then sending them into the model for classification. However, this shallow concatenation has three main drawbacks. First, it neglects the fact that speakers are always changing in turn as a conversation progresses. Second, it weakens the relationships between the context utterances as they are organized in the chronological order. Third, due to the maximum sequence length limit (e.g., 512 for BERT-Base), pre-trained models are unable to tackle sequences that are composed of thousands of tokens, which, however, is a typical setup in multi-turn conversations.

In this paper, we attempt to employ the pre-trained language model and adjust it to fit the task of multi-turn response selection, in which BERT [1] is adopted as the basis of our work. We propose a new model, named Speaker-Aware BERT (SA-BERT). First, to make the pre-trained language model aware of the speaker change information during the conversation, the model is enhanced by adding the speaker embeddings to the token representation and adding the special segmentation tokens between the context utterances. These two strategies are designed to improve the conversation understanding capability. Furthermore, to tackle the entangled dialogues which are mixed with multiple conversation topics and are composed of hundreds of utterances, we propose a heuristic speaker-aware disentanglement strategy, which helps to select a small number of most important utterances according to the speaker change information. Finally, domain adaptation is designed to incorporate specific in-domain knowledge into pre-trained language models.

We perform the adaptation process with the same domain but different sets under the same setting. We can conclude that adaptation on a domain-specific corpus can help to incorporate more domain-specific knowledge, and the more similar to the task this adaptation corpus is, the more improvement it can help to achieve.

We test our model on five datasets, Ubuntu Dialogue Corpus V1 [6], Ubuntu Dialogue Corpus V2 [7], Douban Conversation Corpus [10], E-commerce Dialogue Corpus [12], and DSTC 8-Track
2-Subtask 2 Corpus [5]. Experimental results show that the proposed model outperforms the existing models on all metrics by large margins. Specifically, 5.5% R@10 on Ubuntu Dialogue Corpus V1, 5.9% R@10 on Ubuntu Dialogue Corpus V2, 3.2% MAP and 2.7% MRR on Douban Conversation Corpus, 8.3% R@10 on E-commerce Corpus, and 15.5% R@10 on DSTC 8-Track 2-Subtask 2 Corpus, leading to new state-of-the-art performances for multi-turn response selection.

In summary, our contributions in this paper are three-fold:

1) A new model, named Speaker-Aware BERT (SA-BERT), is designed by employing speaker embeddings and speaker-aware disentanglement strategy, to make BERT aware of the speaker change information as the conversation progresses.
2) We make further analysis on the effect of adaptation to the performance of response selection.
3) Experimental results show that our model achieves new state-of-the-art performances for multi-turn response selection.

2 METHODOLOGY

Given a dialogue dataset \(D\), an example of the dataset is denoted as \((c, r, y)\), where \(c = \{u_1, u_2, ..., u_n\}\) represents a context with \(\{u_k\}_{k=1}^{n}\) as the utterances, \(r\) is a response candidate, and \(y \in \{0, 1\}\) denotes a label. Specifically, \(y = 1\) indicates that \(r\) is a proper response for \(c\); otherwise \(y = 0\). Our goal is to learn a matching model \(g(c, r)\) by minimizing a cross-entropy loss function from \(D\). For any \((c, r)\) pair, \(g(c, r)\) measures the matching degree between \(c\) and \(r\).

We present here our proposed model, named Speaker-Aware BERT (SA-BERT). Readers can refer to [1] for details of BERT.

2.1 Speaker Embeddings & Segmentations

In order to distinguish utterances in a context and model the speaker change in turn as the conversation progresses, we use two strategies to construct the input sequence for multi-turn response selection.

First, in order to model the speaker change, we propose to add additional speaker embeddings to token representations. The embedding functions as indicating the speaker’s identity for each utterance. For conversations with two speakers, two speaker embedding vectors need to be estimated during the training process. The first vector is added to each token of utterances of the first speaker. When the speaker changes, the second vector is employed. This is done by minimizing a cross-entropy loss function from \(D\).

Second, empirical results in [2] show that segmentation tokens play an important role for multi-turn response selection. To model conversation, it is natural to extend that to further model turns and utterances. In this work we propose and empirically show that using an [EOU] token at the end of an utterance and an [EOT] token at the end of a turn model interactions between utterances in a context implicitly and improve the performance consistently.

2.2 Speaker-Aware Disentanglement Strategy

When more than two speakers are communicating in a common channel, there are often multiple conversation topics occurring concurrently. In terms of a specific conversation topic, utterances relevant to it are useful and other utterances could be considered as noise for them. Note that BERT is not good at dealing with sequences which are composed of more tokens than the limit (i.e., length of time steps is set to be 512). In order to select a small number of most important utterances, in this paper, we propose a heuristic speaker-aware disentanglement strategy as follows.

First, we define the speaker who is uttering an utterance as the spoken-from speaker, and define the speaker who is receiving an utterance as the spoken-to speaker. Each utterance usually has the labels of both spoken-from and spoken-to speakers, which can be extracted from the utterance itself. But some utterances may have only the spoken-from speaker label while the spoken-to speaker is unknown which is set to None in our experiments. Second, given the spoken-from speaker of the response, we select the utterances which have the same spoken-from or spoken-to speaker as the spoken-from speaker of the response. Third, these selected utterances are then organized in their original chronological order and used to form the filtered context. Finally, the utterances selected according to their spoken-from or spoken-to speaker labels are assigned with the two speaker embedding vectors respectively.

2.3 Domain Adaptation

The original BERT is trained on a large text corpus to learn general language representations. To incorporate specific in-domain knowledge, adaptation on in-domain corpora are designed. In our experiments, we employ the training set of each dataset for domain adaptation without additional external knowledge. Furthermore, domain adaptation is done by performing multi-task learning that optimizing a combination of two loss functions: next sentence prediction (NSP) and masked language model (MLM) losses [1]. Specifically, the speaker embeddings can be pre-trained in the task of NSP. If there is no any adaptation processes, the speaker embeddings have to be initialized randomly at the beginning of the fine-tuning.

2.4 Output Representation

The first token of each concatenated sequence is the [CLS] token, with its embedding being used as the aggregated representation for a context-response pair classification. This embedding captures the matching information between a context-response pair, which is sent into a classifier with a sigmoid output layer. Finally, the classifier returns a score to denote the matching degree of this pair.

| Dataset       | Train | Valid | Test  |
|---------------|-------|-------|-------|
| Ubuntu V1     | 1M    | 0.5M  | 0.5M  |
| Ubuntu V2     | 1M    | 195k  | 189k  |
| Douban        | 1M    | 50k   | 10k   |
| E-commerce    | 1M    | 10k   | 10k   |
| DSTC 8        | 11M   | 1M    | 1M    |

Table 1: Statistics of the datasets that our model is tested on.
We tested SA-BERT on five public multi-turn response selection datasets, Ubuntu Dialogue Corpus V1 [6], Ubuntu Dialogue Corpus V2 [7], Douban Conversation Corpus [10], E-commerce Dialogue Corpus [12] and DSTC 8-Track 2-Subtask 2 Corpus [5]. The first four datasets have been disentangled in advance by their publishers and our proposed speaker-aware disentanglement strategy is applied to only the last DSTC 8-Track 2-Subtask 2 Corpus. Some statistics of these datasets are provided in Table 1.

### 3 EXPERIMENTS

#### 3.1 Datasets

We tested SA-BERT on five public multi-turn response selection datasets, Ubuntu Dialogue Corpus V1 [6], Ubuntu Dialogue Corpus V2 [7], Douban Conversation Corpus [10], E-commerce Dialogue Corpus [12] and DSTC 8-Track 2-Subtask 2 Corpus [5]. The first four datasets have been disentangled in advance by their publishers and our proposed speaker-aware disentanglement strategy is applied to only the last DSTC 8-Track 2-Subtask 2 Corpus. Some statistics of these datasets are provided in Table 1.

#### 3.2 Evaluation Metrics

We used the same evaluation metrics as those used in previous work [5–7, 10, 12]. Each model was tasked with selecting the $k$ best-matched responses from $n$ available candidates for the given conversation context $c$, and we calculated the recall of the true positive replies among the $k$ selected responses, denoted as $R_n@k$. In addition to $R_n@k$, we considered mean average precision (MAP), mean reciprocal rank (MRR) and precision-at-one (P@1), especially for the Douban corpus, following settings of previous work.

#### 3.3 Experimental Results

Table 2, Table 3 and Table 4 present the evaluation results of SA-BERT and previous methods on the five datasets. All the results except ours are from the existing literature. Due to previous methods did not make use of pre-trained language models, we reproduced the results of BERT baseline by fine-tuning on the training set for reference, denoted as BERT for fair comparisons. As we can see that, BERT has already outperformed the present models on most metrics, except $R_{10}@5$ on Ubuntu Dialogue Corpus V1 and $R_{10}@1$ on E-commerce Corpus. Furthermore, SA-BERT outperformed the present state-of-the-art performance by large margins of 5.5% $R_{10}@1$ on Ubuntu Dialogue Corpus V1, 5.9% $R_{10}@1$ on Ubuntu Dialogue Corpus V2, 3.2% MAP and 2.7% MRR on Douban Conversation Corpus, 8.3% $R_{10}@1$ on E-commerce Corpus, and 15.5% $R_{20}@1$ on DSTC 8-Track 2-Subtask 2 Corpus. These results show the ability of SA-BERT to select the best-matched response.
and its compatibility across domains (system troubleshooting, social network and e-commerce), achieving a new state-of-the-art performance for multi-turn response selection. Our code has been published to help replicate our results\(^1\).

4 ANALYSIS

4.1 Adaptation Corpus

Table 5: Results on the test set of Ubuntu Corpus V2, by adapting domain with different corpora and fine-tuning all on the training set of Ubuntu Corpus V2.

| Corpus         | R\(_@1\) | R\(_@10\) | R\(_@10@2\) | R\(_@10@5\) |
|----------------|---------|---------|-------------|-------------|
| None           | 0.950   | 0.786   | 0.890       | 0.981       |
| DSTC8          | 0.954   | 0.803   | 0.902       | 0.981       |
| Ubuntu V1      | 0.961   | 0.824   | 0.914       | 0.985       |
| Ubuntu V2      | 0.963   | 0.830   | 0.919       | 0.985       |

We make some further analysis on the effect of adaptation corpus to the performance of multi-turn response selection. We performed the adaptation process with the same domain but different sets. Here, three different sets of Ubuntu were employed: DSTC 8-Track 2, Ubuntu Dialogue Corpus V1, and Ubuntu Dialogue Corpus V2. And then the fine-tuning process was all performed on the training set of Ubuntu Dialogue Corpus V2. The results on the test set of Ubuntu Dialogue Corpus V2 were shown in Table 5.

As we can see that, domain adaptation helps to improve the performance no matter which adaptation corpus was used. Furthermore, adaptation and fine-tuning on the same corpus achieved the best performance. One explanation may be that although pre-trained language models are designed to provide general linguistic knowledge, some domain-specific knowledge is also necessary for a specific task. Thus, adaptation on a domain-specific corpus can help to incorporate more domain-specific knowledge. The more similar to the task this adaptation corpus is, the more improvement it can help to achieve.

4.2 Speaker Embeddings

Table 6: Results on the test set of Ubuntu Corpus V2, by ablating the speaker embeddings (SE).

| Pre-Train | SE  | R\(_@1\) | R\(_@10\) | R\(_@10@2\) | R\(_@10@5\) |
|-----------|-----|---------|---------|-------------|-------------|
| No        | No  | 0.950   | 0.781   | 0.890       | 0.980       |
| No        | Yes | 0.950   | 0.786   | 0.890       | 0.981       |
| Yes       | No  | 0.961   | 0.825   | 0.915       | 0.984       |
| Yes       | Yes | 0.963   | 0.830   | 0.919       | 0.985       |

The speaker embeddings were ablated and the results were reported in Table 6. The first two lines discussed the situation in which the adaptation process were omitted, and the last two lines discussed the adaptation process were equipped with. The performance drop verified the effectiveness of speaker embeddings.

4.3 Speaker-Aware Disentanglement Strategy

To show the effectiveness of the speaker-aware disentanglement strategy, we also applied it to the existing model, such as IMN [3]. The original IMN did not employ any disentanglement strategy and selected the last 70 utterances as the context, which achieved a performance of 32.2\% R\(_@10@1\). After employing the strategy, about 25 utterances were selected to form the context, which achieved a performance of 37.5\% R\(_@10@1\). Similar results can also be observed by employing this strategy to BERT and ablating this strategy in SA-BERT, as shown in Table 4, which verified the effectiveness of the speaker-aware disentanglement strategy again.

5 CONCLUSION

In this paper, we study the problem of employing pre-trained language models for multi-turn response selection in retrieval-based chatbots. A speaker-aware model and a speaker-aware disentanglement strategy are proposed. Experiments on five public datasets show that our proposed method achieves a new state-of-the-art performance for multi-turn response selection. Adjusting pre-trained language models to fit multi-turn response selection and designing new disentanglement strategies will be a part of our future work.

REFERENCES

[1] 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 NAACL-HLT. 4171–4186.
[2] Jianxiong Dong and Jim Huang. 2018. Enhance word representation for out-of-vocabulary on Ubuntu dialogue corpus. CoRR abs/1802.02614 (2018).
[3] Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive Matching Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In Proceedings of the 28th ACM International Conference of CIKM 2019. 2321–2324.
[4] Matthew Henderson, Ivan Vulic, Daniela Gerz, Iñigo Casanueva, Pawel Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrksic, and Pei-Hao Su. 2019. Training Neural Response Selection for Task-Oriented Dialogue Systems. In Proceedings of the 57th Conference of ACL 2019. 5392–5404.
[5] Seokhwan Kim, Micel Galle, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, et al. 2019. The Eighth Dialo System Technology Challenge. arXiv preprint arXiv:1911.06394 (2019).
[6] 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 Proceedings of the SIGDIAL 2015 Conference. 285–294.
[7] Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training End-to-End Dialogue Systems with the Ubuntu Dialogue Corpus. (2017). 31–65.
[8] 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 Proceedings of the Twelfth ACM International Conference of WSDM 2019. 267–275.
[9] 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 Proceedings of the 57th Conference of ACL 2019. 1–11.
[10] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhusong Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In Proceedings of the 55th Conference of ACL 2017. 496–505.
[11] Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jiu Hong Han, and Songlin Hu. 2019. Multi-hop Selector Network for Multi-turn Response Selection in Retrieval-based Chatbots. In Proceedings of the 2019 Conference of EMNLP-JKCEPL. 111–120.
[12] Zhaosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. In Proceedings of the 27th International Conference of COLING 2018. 3740–3752.
[13] 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 Proceedings of the 56th Conference of ACL 2018. 1118–1127.

\(^1\)https://github.com/JasonForJoy/SA-BERT